SCALABLE ANALYSIS AND DESIGN OF SERVICE SYSTEMS

A Thesis
Presented to
The Academic Faculty

by

Bo Zhang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
May 2011

SCALABLE ANALYSIS AND DESIGN OF SERVICE SYSTEMS

Approved by:

Professor Albertus P. Zwart,
Committee Chair
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Hayriye Ayhan
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Jiangang Dai
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Robert Foley
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Dr. Martin Reiman
Bell Labs
Alcatel-Lucent

Date Approved: 21 March 2011

To my parents...

ACKNOWLEDGEMENTS

My research described in this dissertation is partly supported by NSF grants CMMI-0727400 and CMMI-1030589. Part of the research was carried out during my visit/internship at Bell Labs in New Jersey, Centrum Wiskunde and Informatica in Amsterdam, and NYU Stern School of Business in New York. Their hospitality is gratefully acknowledged.

For this thesis and for my intellectual and professional development as a Ph.D. student, I am most indebted to my advisor Bert Zwart. All the work described here, except Section 3.7, has been guided by Bert. I was truly lucky to have a scholar of his calibre to teach me how to do research, in particular how to do math. This dissertation and my Ph.D. training have benefited in many ways from his professionalism as a researcher and as a mentor. His words of encouragement will continue inspiring me in the years to come and I will always be grateful for everything he has done for me.

Chapter III is also joint work with Johan van Leeuwaarden, who has generously shared with me his expertise in asymptotics and his talent for presenting mathematical research. I would also like to thank Johan for the support that he has showed for me as a junior colleague of his. Section 3.7.1 is a product of the collaboration with Josh Reed. Our collaboration has been a great learning experience for me due to his intelligence.

I would also like to thank the four journal referees who have reviewed the two papers corresponding to Chapters II and III for their helpful comments.

It is my pleasure to have Hayriye Ayhan, Jim Dai, Bob Foley, and Marty Reiman to serve on my thesis committee. I am thankful for every piece of advice I have received from them.

In addition, for my excellent Ph.D. experience, I would like to first thank Hayriye Ayhan for being my research co-advisor. What I have learned in working with her, from her expertise in Markov decision processes and from her approach to doing research, will always be a valuable tool in my research arsenal. Also, I am thankful for her financial

support during the summer of 2010.

I was blessed to have Sem Borst and Marty Reiman as my mentors during my visit at Bell Labs. From our interactions I observed masterful problem-solving, pure scholarship, as well as genuine humbleness and decency. What I have learned from them will influence me no matter what career I pursue.

I was fortunate to take or audit courses taught by several great professors at ISyE, Professors Sigrun Andradottir, Jim Dai, and Bob Foley, and some excellent teachers in the math department, Professors Dmitriy Bilyk, Chris Heil, Plamen Iliev, and Ionel Popescu, and to attend reading seminars organized by Professor Dai. What I learned in those classes and seminars has been helpful for my research. Also, I would like to thank Professor Gary Parker, our Ph.D. program director, for his support for me as a Ph.D. student.

Many people helped me before I came to Georgia Tech. I am especially grateful to Professors Sheldon Ross and Maged Dessouky at USC. Professor Ross introduced me to probability with his legendary lectures and textbooks. Our interactions have been influential to my way of thinking. Professor Dessouky generously provided me with a research assistantship opportunity. I am grateful to him for his guidance on my first real research project and the trust and support that he showed for me.

My friends have shaped me into the person I am. I would like to thank them for their help, encouragement, and support in my academic pursuit.

Finally, I owe my deepest gratitude to my parents and my wife. None of this would have been possible without their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In this dissertation, we develop analytical and computational tools for performance analysis and design of large-scale service systems. The dissertation consists of three main chapters.

The first chapter is devoted to devising efficient task assignment policies for large-scale service system models from a rare event analysis standpoint. Specifically, we study the steady-state behavior of multi-server queues with general job size distributions under size-interval task assignment (SITA) policies. Assuming Poisson arrivals and the existence of the αth moment of the job size distribution for some $\alpha > 1$, we show that if the job arrival rate and the number of servers increase to infinity with the traffic intensity held fixed, a SITA policy parameterized by α minimizes in a large deviation sense the steady-state probability that the total number of jobs in the system is greater than or equal to the number of servers. The optimal large deviation decay rate can be arbitrarily close to the one for the corresponding probability in an infinite-server queue, which only depends on the system traffic intensity but not on any higher moments of the job size distribution. This supports in a many-server asymptotic framework the common wisdom that separating large jobs from small jobs protects system performance against job size variability.

In the second chapter, we study constraint satisfaction problems for a Markovian parallel-server queueing model with impatient customers, motivated by large telephone call centers. To minimize the staffing level subject to different service-level constraints, we propose refined square-root staffing (SRS) rules, which preserve the insightfulness and computational scalability of the celebrated SRS principle and yet achieve a stronger form of optimality. In particular, using asymptotic series expansion techniques, we first develop refinements to a set of asymptotic performance approximations recently used in analyzing large call centers, namely, the Quality and Efficiency Driven (QED) diffusion approximations. We then use the improved performance approximations to explicitly characterize the error of

conventional SRS and further obtain the refined SRS rules. Finally, we demonstrate how the explicit form of the staffing refinements enables an analytical assessment of the accuracy of conventional SRS and its underlying QED approximation.

In the third chapter, we study a fluid model for many-server Markovian queues in changing environments, which can be used to model large-scale service systems with customer abandonments and time-varying arrivals. We obtain the stationary distribution of the fluid model, which refines and is shown to converge, as the environment changing rate vanishes in a proper way, to a simple discrete bimodal approximation. We also prove that the fluid model arises as a law of large number limit in a many-server asymptotic regime.

CHAPTER I

INTRODUCTION

In this dissertation, we develop analytical and computational tools for performance analysis and design of large-scale service systems. The dissertation consists of three main chapters.

The first chapter is devoted to devising efficient task assignment policies for large-scale service system models from a rare event analysis standpoint. Specifically, we study the steady-state behavior of multi-server queues with general job size distributions under size-interval task assignment (SITA) policies. Assuming Poisson arrivals and the existence of the αth moment of the job size distribution for some $\alpha > 1$, we show that if the job arrival rate and the number of servers increase to infinity with the traffic intensity held fixed, a SITA policy parameterized by α minimizes in a large deviation sense the steady-state probability that the total number of jobs in the system is greater than or equal to the number of servers. The optimal large deviation decay rate can be arbitrarily close to the one for the corresponding probability in an infinite-server queue, which only depends on the system traffic intensity but not on any higher moments of the job size distribution. This supports in a many-server asymptotic framework the common wisdom that separating large jobs from small jobs protects system performance against job size variability.

In the second chapter, we study constraint satisfaction problems for a Markovian parallel-server queueing model with impatient customers, motivated by large telephone call centers. To minimize the staffing level subject to different service-level constraints, we propose refined square-root staffing (SRS) rules, which preserve the insightfulness and computational scalability of the celebrated SRS principle and yet achieve a stronger form of optimality. In particular, using asymptotic series expansion techniques, we first develop refinements to a set of asymptotic performance approximations recently used in analyzing large call centers, namely, the Quality and Efficiency Driven (QED) diffusion approximations. We then use the improved performance approximations to explicitly characterize the error of

conventional SRS and further obtain the refined SRS rules. Finally, we demonstrate how the explicit form of the staffing refinements enables an analytical assessment of the accuracy of conventional SRS and its underlying QED approximation.

In the third chapter, we study a fluid model for many-server Markovian queues in changing environments, which can be used to model large-scale service systems with customer abandonments and time-varying arrivals. We obtain the stationary distribution of the fluid model, which refines and is shown to converge, as the environment changing rate vanishes in a proper way, to a simple discrete bimodal approximation. We also prove that the fluid model arises as a law of large number limit in a many-server asymptotic regime.

CHAPTER II

STEADY-STATE ANALYSIS FOR MULTI-SERVER QUEUES UNDER SIZE-INTERVAL TASK ASSIGNMENT IN THE QUALITY-DRIVEN REGIME

We study the steady-state behavior of multi-server queues with general job size distributions under size-interval task assignment (SITA) policies. Assuming Poisson arrivals and the existence of the αth moment of the job size distribution for some $\alpha > 1$, we show that if the job arrival rate and the number of servers increase to infinity with the traffic intensity held fixed, a SITA policy parameterized by α minimizes in a large deviation sense the steady-state probability that the total number of jobs in the system is greater than or equal to the number of servers. The optimal large deviation decay rate can be arbitrarily close to the one for the corresponding probability in an infinite-server queue, which only depends on the system traffic intensity but not on any higher moments of the job size distribution. This supports in a many-server asymptotic framework the common wisdom that separating large jobs from small jobs protects system performance against job size variability.

2.1 Introduction

Multi-server queues with Poisson arrivals and general job size distributions are probably the simplest, yet reasonable models for many real-world service systems, such as server farms and call centers. One of the oldest and most fundamental problems arising in this type of systems is the choice of a good rule for assigning jobs (or tasks) to servers, known as the *task assignment policy*. In general, it is rather difficult to evaluate the performance of almost any task assignment policy in such systems, and thus many basic questions remain open on this subject.

One intriguing open question that motivates our study is regarding the performance of size-based task assignment, which broadly refers to the practice of dispatching jobs of different sizes or different size distributions to separate server pools. The question is whether or

not separating large jobs from small ones protects system performance against high variability of job sizes. Intuitively, such a task assignment policy prevents large jobs from occupying all servers and blocking many small jobs. This advantage should become especially apparent as the job size variability increases. On the other hand, dedicating servers to large and small jobs respectively can lead to server idling and underutilization, and thus hurt system performance. This suggests that the right answer to the proposed question should depend on the context, and the specifics of the policy matter.

In this chapter, we focus on one type of size-based task assignment policy, namely, the size-interval task assignment (SITA) policy. First formally introduced by Harchol-Balter et al. [26], the SITA policy has since attracted a lot of research attention (*e.g.*, [3, 12, 14, 27, 28, 46]), especially in the computer systems performance evaluation community (see further discussion later in this section). In words, a SITA policy groups all servers into multiple pools, divides the possible job sizes into different intervals, and assigns all jobs in each size interval to one server pool exclusively, with each subsystem operating on an FCFS basis. In Section 2.2, we shall provide a more precise definition of SITA.

Our goal is to study, for a system where SITA has been chosen as the scheduling discipline a priori, how to determine the parameters of the SITA policy in order to minimize the occurence of congestion, when the number of servers is large and the traffic intensity in the system is bounded away from 1. Specifically, we study the steady-state performance of the proposed SITA policies in the Quality-Driven (QD) regime, where the arrival rate and the number of servers N increase to infinity with the traffic intensity (<1) held fixed. Loosely speaking, the QD regime is a many-server, overstaffing regime with a constant overstaffing factor, and it is appropriate for modeling large-scale systems with a non-negligible amount of slack capacity. The practice of reserving extra capacity is often seen in service-oriented systems, such as emergency call centers, and a variety of computer systems where capacity over-provisioning is commonplace to meet high quality-of-service standards (see [37] and the references therein).

The performance measure that we are interested in is $\mathrm{P}\{Q^N \geq N\}$, in particular its rate of decay in the QD regime, where Q^N denotes the steady-state total number of jobs in the

N-server system. This probability is a good indication of system performance, especially in the QD regime. First, the occurrence of the event {number of jobs in the system $\geq$ number of servers} implies that there are jobs being delayed. In practice, this could suggest a greater chance of hardware failure or incur extra operational cost due to using spare or outsourced capacity (so-called pay-per-use; see [36]). In addition, in the QD regime, $\mathrm{P}\{Q^N \geq N\}$ should become a very remote tail probability as N increases to infinity, which makes it a natural quantity of interest from the point of view of large deviations theory and extreme value theory (see [16, 17]). Furthermore, we believe that studying the asymptotics of $\mathrm{P}\{Q^N \geq N\}$ is an important step towards a complete understanding of the distribution of Q^N and other steady-state performance metrics.

The performance benchmark that we use is the lower bound $\mathrm{P}\{Q_\infty^N \geq N\}$, where Q_∞^N denotes the steady-state number of jobs in the corresponding M/G/∞ queue. In other words, we are interested in how close to $\mathrm{P}\{Q_\infty^N \geq N\}$ a cleverly chosen SITA policy can make $\mathrm{P}\{Q^N \geq N\}$. The M/G/∞ queue is a remarkably tractable model and has been exploited to analyze the M/G/N queue in other regimes (*e.g.*, see [55]).

Next, we provide a brief review of the relevant literature. As mentioned earlier, SITA has been extensively studied by the computer systems performance evaluation community. In Section 3 of [26], Harchol-Balter et al. state that multi-server queues under SITA were inspired by and used as an abstraction of the `xolas` distributed computing facility at MIT's Laboratory for Computer Science. They also argue that the assumption that task sizes are known holds to an approximate degree in other contexts, such as some batch computing servers. Ciardo et al. [14] apply SITA to web server farms, and they point out that if the exact size of each job (*i.e.*, URL request in that setting) is not available to the front-end dispatcher, SITA still can be implemeted by a two-stage allocation policy. Schroeder and Harchol-Balter [46] apply SITA to heavy-tailed supercomputing workloads and they argue that in many distributed servers task assignment is done by the users (rather than a dispatcher); specifically, in the SITA case, each job is submitted with an estimated runtime and different host machines have different duration limitations: up to 2 hours, up to 4 hours, up to 8 hours, or unlimited, etc. Cardellini et al. [12] also discuss using SITA for

web server farms, especially when the Web content is static. For a more complete list of references and a careful discussion on the existing results, we refer the reader to Section 2 in [27]. In addition, we note that size-based scheduling policies, such as shortest-job-first, preemptive-shortest-job-first, and shortest-remaining-processing-time-first, are well-studied subjects in the literature of single-server queues (see [25, 41]).

In terms of the performance of SITA or more generally size-based task assignment, both positive and negative results have been reported. These results are often provided through a comparison with the work-conserving First-Come-First-Served (FCFS) policy, *i.e.*, the policy of forming one single queue in front of the whole server pool and processing jobs on an FCFS basis. For example, Smith and Whitt [48] demonstrate that as job size variability increases, the steady-state queue length of a Poisson input two-server queue with a hyperexponential job size can go to infinity under the FCFS policy, while the same system under a size-based task assignment policy that essentially divides the system into two independent M/M/1 queues maintains a constant steady-state queue length independent of the variability level. On the other hand, a recent study by Harchol-Balter et al. [27] shows that the mean job delay in a two-server queue under a size-based policy can diverge, while it converges under FCFS, as the variance of the job size distribution goes to infinity. In fact, in [27] Harchol-Balter et al. show that neither size-based task assignment nor FCFS could be a sure win.

To the best of our knowledge, a proof on the superiority (or inferiority) of sized-based task assignment policies has never materialized in any general setting (see [27]). This lack of analytical results is largely due to the fact that the study of multi-server queues under sized-based task assignment boils down to analyzing several M/G/N subsystems under FCFS in parallel, while the M/G/N queue under FCFS remains an unsolved problem (see [34]). Most existing results on M/G/N are approximate; for example, see [53, 54, 61] and the references therein. In the many-server asymptotic setting, the only analytical result on steady-state distributions is provided by Gamarnik and Momčilović [18], who obtain an explicit expression for the critical exponent for the moment generating function of a limiting (scaled) steady-state queue length in the Quality-and-Efficiency-Driven regime assuming

lattice-valued job sizes with a finite support (and thus with a light tail). As for steady-state analysis for many-server queues with heavy-tailed job sizes, we are not aware of any conclusive result. Whitt [55] shows that the steady-state waiting-time distribution of the M/G/N queue under FCFS has a heavy tail (with appropriate definition) whenever the job size distribution does. Also, to the best of our knowledge, task assignment has never been studied in a many-server regime.

The QD regime has not received much attention in the literature. The service quality in the QD regime is noted to be extremely good for a large number of servers N (see [21, 42, 62]), hence the name. However, it seems that there has almost been no study that attempts to quantify how fast the performance improves as N increases, which is essential for resource dimensioning. The only work that we are aware of is [62], where the performance asymptotics for the M/M/N+G queue in the QD regime are derived based on the exact formulas. Also, we note that estimating the performance of many-server queues by simulation is difficult as well, because it often involves increasingly rare events as N grows. A recent study in this regard is the paper by Blanchet et al. [6], which develops a rare-event simulation algorithm that is asymptotically efficient in the QD regime. We finally note that there does exist an extensive literature on the analysis of queues where there is a finite number of servers, of which the speed grows, and the number of input processes grows accordingly. This is also known as the many-flows regime, and is motivated by highly multiplexed communication networks; see [19] for background.

In short, the main contribution of this chapter is the construction of a family of SITA policies under which $\mathrm{P}\{Q^N \geq N\}$ can be made arbitrarily close to $\mathrm{P}\{Q^N_\infty \geq N\}$ in a large deviation sense. Our result holds true for any job size distribution with finite αth moment for some $\alpha > 1$, including those with an infinite variance or a heavy-tailed distribution. Furthermore, if the job size has a finite support (*i.e.*, finitely many possible values) or a hyperexponential distribution, our proposed size-based task assignment policy achieves an even stronger performance: $\mathrm{P}\{Q^N \geq N\} \sim \mathrm{P}\{Q^N_\infty \geq N\}$ on a logarithmic scale. Because $\mathrm{P}\{Q^N_\infty \geq N\}$ is insensitive to the job size distribution, our results suggest that the recommended size-based task assignment policies indeed protect system performance against

high job size variability in the QD regime. In the general job size distribution case, the number of size intervals (or equivalently, server pools) in our SITA prescription grows to infinity at sublinear rate as N increases, and both the size cutoff values and the number of servers allocated to each pool are carefully parameterized by the job size moment index α. These features enable our policies to perform well even when the job size has an unbounded support. Our approach to analyzing systems with many servers and job sizes of unbounded support is potentially interesting for other problems, such as the behavior of the same probability under FCFS in the QD regime, which seems open. In obtaining the main results, we also develop some estimation results on random walks and the M/D/N queue, both asymptotics and bounds, which may be of independent interest. In particular, there is a connection with light traffic analysis of random walks (see [2]).

In the next section, we formulate our model, state our main results in more detail, and provide the organization of the remainder of this chapter.

2.2 Model formulation and main results

In this section, we first provide the model formulation, including the definitions of the QD limiting regime and the SITA policy. Then, we summarize the main results of this chapter.

We study the M/G/N queue in the QD regime, which is achieved by considering a sequence of queues indexed by the number of servers N where the arrival rate to the system grows large proportionally to N and the traffic intensity remains fixed. Specifically, first let $\{A(t), t \geq 0\}$ be a Poisson process with rate λ, and T_i, $i \geq 1$, be the interarrival times. In the Nth system, let the job arrival process, namely, $\{A^N(t), t \geq 0\}$, be such that $A^N(t) = A(Nt)$ for all $t \geq 0$, or equivalently the interarrival times T_i^N, $i \geq 1$, are equal to T_i/N. Denoting by λ^N the arrival rate to the Nth system, we then have $\lambda^N = N\lambda$. Both the job size distribution and the traffic intensity are the same for any N. Specifically, job sizes are i.i.d. (independent and identically distributed), equal in distribution to S, with $\mathrm{E}[S] = \mu^{-1}$, and each server processes jobs at unit rate. As a result, the traffic intensity in the Nth system is just $\lambda^N/N\mu = \lambda/\mu$, and we denote this fixed traffic intensity by $\rho := \lambda/\mu$. The QD regime is achieved by letting $N \to \infty$. We further assume $\mathrm{P}\{S > 0\} = 1$ for simplicity. This

assumption is not restrictive, because if $\mathrm{P}\{S > 0\} < 1$, one may just allocate one server for zero-size jobs and apply our proposed policies, with N replaced by $N-1$, to the other servers and jobs, leading to the same optimality results established in this chapter.

As a performance benchmark for the sequence of multi-server queues in the QD regime, we also consider a sequence of infinite-server queues: in the Nth system, jobs with i.i.d. sizes equal in distribution to S arrive according to the process $\{A^N(t), t \geq 0\}$ to an infinite number of servers and are served immediately upon arrival. We denote by Q^N_∞ the steady-state number of jobs in the Nth infinite-server queue.

Next, we provide a mathematical definition of a SITA policy prescription. Let $\mathbb{N}$ denote the set of natural numbers, $\mathbb{R}_+ = [0, \infty)$, $\mathbb{Z}_+ = \mathbb{N} \cup \{0\}$. A SITA policy prescription, say, denoted by $\pi(N)$, is fully characterized by the following parameters:

- a positive-integer-valued function $m(N)$,
- an $[m(N)+1]$-dimensional function $\{r_i(N), i = 0, ..., m(N)\} \in \mathbb{R}_+^{m(N)+1}$, which is increasing in i,
- an $m(N)$-dimensional function $\{s_i(N), i = 1, ..., m(N)\} \in \mathbb{Z}_+^{m(N)}$ satisfying $\sum_{i=1}^{m(N)} s_i(N) = N$.

Specifically, in an N-server system under policy $\pi(N)$, jobs are divided into $m(N)$ types according to their sizes: all jobs with their size in the interval $(r_{i-1}(N), r_i(N)]$ are type i jobs, with $r_0(N) \equiv 0$ and $r_{m(N)}(N)$ equal to the largest possible value of the job size (which might equal infinity). $s_i(N)$ of the N servers form a pool processing type i jobs exclusively on an FCFS basis.

To state our main results, we start by defining two notions of optimality with respect to the rate at which $\mathrm{P}\{Q^N \geq N\}$ decays to zero in the QD regime, compared to the benchmark $\mathrm{P}\{Q^N_\infty \geq N\}$.

Definition 2.2.1. *A policy $\pi(N)$ is strongly optimal in the QD regime, if the steady-state total number of jobs in the system under $\pi(N)$, denoted by Q^N, satisfies*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q^N \geq N\} = \lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q^N_\infty \geq N\}. \tag{1}$$

Definition 2.2.2. *A family of policies $\{\pi_\epsilon(N)\}_{\epsilon>0}$ are weakly optimal in the QD regime, if for all $\epsilon > 0$, the steady-state total number of jobs in the system under $\pi_\epsilon(N)$, denoted by Q_ϵ^N, satisfies*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q_\epsilon^N \geq N\} - \lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q_\infty^N \geq N\} < \epsilon. \tag{2}$$

These two definitions make sense because, as we shall show in Section 3.1,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q_\infty^N \geq N\} \in (-\infty, 0) \tag{3}$$

and

$$\mathrm{P}\{Q_\infty^N \geq N\} \leq \mathrm{P}\{Q^N \geq N\} \quad \text{for all } N. \tag{4}$$

Also due to (4), inequality (2) implies that if $\lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q_\epsilon^N \geq N\}$ exists, then

$$\lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q_\infty^N \geq N\} \leq \lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q_\epsilon^N \geq N\} < \lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q_\infty^N \geq N\} + \epsilon, \tag{5}$$

which means that the large deviation decay rate of $\mathrm{P}\{Q_\epsilon^N \geq N\}$ is ϵ-close to that of $\mathrm{P}\{Q_\infty^N \geq N\}$.

Our first main result states that, if the job size S can take on, say, m possible values, the policy of dividing servers into m pools each of which specializes in processing jobs with one of the m possible sizes is strongly optimal, if the workloads are balanced among the m server pools or each pool has (asymptotically) the same traffic intensity as that for the whole system, *i.e.*, ρ. Specifically, suppose each job size equals d_i with probability p_i, $i = 1, ..., m$, where $\sum_{i=1}^m p_i = 1$, $\sum_{i=1}^m p_i d_i = \mu^{-1}$, $2 \leq m < \infty$, and, without loss of generality, $d_1 < d_2 < ... < d_{m-1} < d_m$. We define a SITA policy $\pi(N)$ as follows:

$$m(N) := m, \quad r_0(N) := 0, \quad r_i(N) := d_i, \quad i = 1, ..., m, \tag{6}$$

$$s_i(N) := \lfloor N p_i d_i \mu \rfloor, \quad i = 1, ..., m-1, \qquad s_m(N) = N - \sum_{i=1}^{m-1} s_i(N). \tag{7}$$

This policy is load-balancing, as it follows from (6) and (7) that the traffic intensity to each server pool in the Nth system, namely $\rho_i(N)$, reads

$$\rho_i(N) = \rho \cdot \frac{N p_i d_i \mu}{\lfloor N p_i d_i \mu \rfloor} \approx \rho, \qquad \text{for } i = 1, ..., m-1, \tag{8}$$

$$\rho_m(N) = \rho \cdot \frac{N - \sum_{i=1}^{m-1} N p_i d_i \mu}{N - \sum_{i=1}^{m-1} \lfloor N p_i d_i \mu \rfloor} \approx \rho. \tag{9}$$

This load-balancing SITA policy achieves strong optimality in the QD regime.

Theorem 2.2.3. *Suppose the job size S can take on m different possible values: $d_1 < d_2 < ... < d_{m-1} < d_m$. The SITA policy $\pi(N)$, defined by* (6) *and* (7)*, is strongly optimal in the QD regime.*

This result can be extended to systems with hyperexponentially distributed S with m branches, which are appropriate models for many kinds of service systems, where m types of services are provided and, conditioning on the type of service that a customer requests, his service time is exponentially distributed. Assuming that the type of service requested by each customer is known to the system scheduler (achieved in call centers, for example, by asking customers to select the type of service they need before assigning them to agents), a similar policy of forming m load-balanced server pools each of which provides one type of service is strongly optimal (see Remark 2.4.2).

Our second main result is an extension of the first result to systems with general job size distributions (see Algorithms 1 and 2, Theorems 2.2.4 and 2.2.6, Corollary 2.2.5). Specifically, we provide a family of SITA policies that are weakly optimal for general job size distributions. Our only assumption on S is that $\mathrm{E}[S^\alpha] < \infty$ for some $\alpha > 1$, and therefore our result holds not only for light-tailed job sizes but also for heavy-tailed ones. Although the intuition of dividing servers into load-balanced pools remains the same with general S, the prescription of the SITA policies is much more involved; in fact, because the job size has an unbounded support while only finitely many servers are available, all the policy parameters need to be chosen very carefully. This is the key to achieving weak optimality.

Roughly speaking, in the general job size case, our proposed policies assign one server to process jobs whose size is above a very large threshold value, and enforce load-balancing SITA among all other servers and jobs. This is far from trivial, both when it comes to the formulation of the policies, as well as the associated analysis, as the number of size intervals we consider grows with the system size N. We next describe the policies in detail and state the optimality results.

First, we assume that S is integer-valued. The approach that we take in this scenario captures the essence of how weak optimality is achieved in the general unbounded-support

case, and we believe that both the policy construction and the proof technique used here may be useful in other contexts. Specifically, suppose that the job size S can take on any $i \in \mathbb{N}$ and $\mathrm{E}[S^\alpha] < \infty$ for some $\alpha > 1$. Define $p_i := \mathrm{P}\{S = i\}$, $i \in \mathbb{N}$. Throughout the chapter, we denote the floor and ceiling functions by $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$, respectively, the natural logarithm by log and define

$$I(\rho) := -\log \rho - 1 + \rho. \tag{10}$$

We shall see that $I(\rho)$ is the decay exponent for $\mathrm{P}\{Q_\infty^N \geq N\}$ (see (23)). For any given $\epsilon > 0$, we propose a SITA policy $\pi_\epsilon(N)$ as follows:

Algorithm 1 $\pi_\epsilon(N)$: SITA for integer-valued S with finite αth moment, for $\alpha > 1$

1: Fix $\eta \in (0,1)$ and

$$\gamma \in \left(0, \min\left\{\eta\alpha^{-1}(\alpha-1)^2, 1\right\}\right). \tag{11}$$

2: Find $\sigma(\epsilon) \in (0, 1-\rho)$ such that $\rho_\epsilon := \rho[1-\sigma(\epsilon)]^{-1}$ satisfies

$$-I(\rho) \leq -I(\rho_\epsilon) < -I(\rho) + \epsilon. \tag{12}$$

3: Let

$$f_i(N) := i, \quad \text{for } i = 1, ..., \lceil N^\eta \rceil, \tag{13}$$

$$f_i(N) := f_{i-1}^\alpha(N) N^{-(i-\lceil N^\eta \rceil)\gamma}, \quad \text{for } i = \lceil N^\eta \rceil + 1, ..., 2\lceil N^\eta \rceil - 1. \tag{14}$$

4: Define $\pi_\epsilon(N)$ as follows

$$m(N) := 2\lceil N^\eta \rceil, \quad r_0(N) := 0, \quad r_{2\lceil N^\eta \rceil}(N) := +\infty \tag{15}$$

$$r_i(N) := \lfloor f_i(N) \rfloor, \quad \text{for } i = 1, ..., 2\lceil N^\eta \rceil - 1, \tag{16}$$

$$s_i(N) := \lceil N P_{L_i} f_i(N) \mu(1-\sigma(\epsilon)) \rceil, \quad \text{for } i = 1, ..., 2\lceil N^\eta \rceil - 2, \tag{17}$$

$$\text{where } P_{L_i} := \mathrm{P}\{S \in (r_{i-1}(N), r_i(N)]\} = \mathrm{P}\{S \in (f_{i-1}(N), f_i(N)]\},$$

$$s_{2\lceil N^\eta \rceil - 1}(N) := N - 1 - \sum_{i=1}^{2\lceil N^\eta \rceil - 2} s_i(N), \quad s_{2\lceil N^\eta \rceil}(N) := 1. \tag{18}$$

The recursive definition (14) is key in the policy prescription. It is significant that

(14) is parameterized by the moment index of the job size distribution α, because this parameterization enables the proposed policy to exploit the information of the job size distribution implied by the moment index.

From (13), (16), and (17), for $i = 1, ..., \lceil N^\eta \rceil$, we easily obtain that

$$r_i(N) = f_i(N) = i \quad \text{and} \quad s_i(N) = \lceil Np_i i\mu(1-\sigma(\epsilon)) \rceil. \tag{19}$$

A simple calculation using (19) further shows that the traffic intensities to the first $\lceil N^\eta \rceil$ server pools are all ρ_ϵ except for a round-off error, just like (7) leading to (8). Hence, under policy $\pi_\epsilon(N)$, the first $\lceil N^\eta \rceil$ subsystems are M/D/$s_i(N)$ queues, all with traffic intensity ρ_ϵ.

For all $i = \lceil N^\eta \rceil + 1, ..., 2\lceil N^\eta \rceil - 1$, we first note that the range of the job sizes in subsystem i is $(r_{i-1}(N), r_i(N)]$ by the definition of SITA; in particular, the maximum possible job size is $r_i(N)$, which is less than or equal to $f_i(N)$ according to (16). In addition, the definition of $s_i(N)$ (17) is exactly specified in such a way that if all jobs in subsystem i had size $f_i(N)$, then this subsystem would also be an M/D/$s_i(N)$ queue with traffic intensity ρ_ϵ (except for a rounding error). In Section 2.5, we shall analyze the performance of these M/D/$s_i(N)$ queues, which serves as an upper bound for the original system.

Finally, (18) states that the last subsystem consists of only one server, reserved for jobs whose sizes are greater than $r_{2\lceil N^\eta \rceil - 1}(N)$. The traffic intensity in this single-server queue turns out to be exponentially small as $N \to \infty$ (see (114)), and therefore the steady-state number of jobs in this single server queue is, loosely speaking, negligible in the QD regime.

In Section 2.5 we shall show that the family of policies $\{\pi_\epsilon(N)\}_{\epsilon>0}$ achieve weak optimality:

Theorem 2.2.4. *Let the job size S be a discrete random variable taking on positive integer values, with $p_i := \mathrm{P}\{S = i\}$, $i \in \mathbb{N}$, and assume $\mathrm{E}[S^\alpha] < \infty$ for some $\alpha > 1$. Then the family of policies $\{\pi_\epsilon(N)\}_{\epsilon>0}$, prescribed by Algorithm 1, are weakly optimal in the QD regime.*

As a consequence of Theorem 2.2.4, if all values in the range of S are divisible by some $\delta > 0$, an easy modification of $\{\pi_\epsilon(N)\}_{\epsilon>0}$ (by measuring time in units of δ's and

thus viewing S as integer-valued) achieves weak optimality. For the sake of brevity, we omit the detailed description of the modified weakly optimal policies. We simply denote by $\{\pi_\epsilon^\delta(N,S)\}_{\epsilon>0}$ the family of SITA policies that are weakly optimal for systems with job size S, where all possible values of S are divisible by δ. Note that by this definition, $\pi_\epsilon(N) = \pi_\epsilon^1(N,S)$.

Corollary 2.2.5. *Let the job size S be a discrete random variable whose possible values are integer multiples of δ, for some $\delta > 0$, and $\mathrm{E}[S^\alpha] < \infty$ for some $\alpha > 1$. The family of policies $\{\pi_\epsilon^\delta(N,S)\}_{\epsilon>0}$ are weakly optimal in the QD regime.*

Finally, we describe the SITA policies for general job size distributions, and state the weak optimality result as Theorem 2.2.6.

Algorithm 2 $\pi_{\epsilon,g}(N)$: SITA for general S with finite αth moment, for $\alpha > 1$

1: Find $\delta_0 = \delta_0(\epsilon) > 0$ such that, with

$$S_{\delta_0} := \delta_0 \left\lceil \frac{S}{\delta_0} \right\rceil \quad \text{and} \quad \rho_{\delta_0} := \lambda \cdot \mathrm{E}[S_{\delta_0}], \tag{20}$$

the following holds:

$$-I(\rho) \le -I(\rho_{\delta_0}) < -I(\rho) + \frac{1}{2}\epsilon. \tag{21}$$

2: Let $\pi_{\epsilon,g}(N) := \pi_{\epsilon/2}^{\delta_0}(N, S_{\delta_0})$ as given in Corollary 2.2.5.

Theorem 2.2.6. *Suppose that there exists $\alpha > 1$ such that $E[S^\alpha] < \infty$. The family of policies $\{\pi_{\epsilon,g}(N)\}_{\epsilon>0}$, prescribed by Algorithm 2, are weakly optimal in the QD regime.*

The remainder of this chapter is devoted to the proofs of the results presented in this section and organized as follows. Section 2.3 contains some preliminary results upon which our proofs of the main theorems will hinge. In Section 2.4, we provide our first main result, namely, the strong optimality of a load-balancing SITA policy for systems where the job size has a finite support. Section 2.5 is devoted to the proof of Theorem 2.2.4, which is the major step towards the construction of weakly optimal policies for general job size distributions. Finally, in Section 2.6, we prove Theorem 2.2.6 and thus establish our second main result. Section 2.8 includes proofs of some technical lemmas.

2.3 Preliminaries

In this section, we develop some preliminary results, which will be useful in proving our main results. These preliminary results may also be of independent interest.

2.3.1 Exact asymptotics for the M/G/∞ queue

This subsection focuses on the exact asymptotic result on $\mathrm{P}\{Q_\infty^N \geq N\}$. For any two real sequences (a^N) and (b^N), we write $a^N \sim b^N$ if $\lim_{N\to\infty}(a^N/b^N) = 1$.

Theorem 2.3.1.

$$\mathrm{P}\{Q_\infty^N \geq N\} \sim \frac{1}{(1-\rho)\sqrt{2\pi N}} \cdot e^{-NI(\rho)}. \tag{22}$$

Therefore,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q_\infty^N \geq N\} = -I(\rho). \tag{23}$$

Proof. First, because Q_∞^N is Poisson distributed with mean $N\rho$ (see (2.40) in [23]), we have that, for all $i \in \mathbb{Z}_+$,

$$\mathrm{P}\{Q_\infty^N = N+i\} = \frac{(N\rho)^i N!}{(N+i)!} \cdot \mathrm{P}\{Q_\infty^N = N\}. \tag{24}$$

It then follows that, for all $i \in \mathbb{Z}_+$,

$$\left(\frac{N\rho}{N+i}\right)^i \cdot \mathrm{P}\{Q_\infty^N = N\} \leq \mathrm{P}\{Q_\infty^N = N+i\} \leq \rho^i \cdot \mathrm{P}\{Q_\infty^N = N\}, \tag{25}$$

and thus

$$\sum_{i=0}^{\infty} \left(\frac{N\rho}{N+i}\right)^i \cdot \mathrm{P}\{Q_\infty^N = N\} \leq \mathrm{P}\{Q_\infty^N \geq N\} \leq \sum_{i=0}^{\infty} \rho^i \cdot \mathrm{P}\{Q_\infty^N = N\}. \tag{26}$$

Applying the Dominated Convergence Theorem to the left-hand side of the first inequality in (26) when letting $N \to \infty$, we obtain that

$$\mathrm{P}\{Q_\infty^N \geq N\} \sim \frac{1}{1-\rho}\, \mathrm{P}\{Q_\infty^N = N\}. \tag{27}$$

The exact asymptotic result (22) follows from (27), Q_∞^N being Poisson with mean ρN, and the Stirling's approximation for the factorial. The logarithmic asymptotic result (23) then follows immediately from (22). $\square$

2.3.2 Exact asymptotics for the M/D/N queue

In this subsection, we consider systems with deterministic job sizes equal to μ^{-1}. The exact asymptotic result that we develop here is new and will be used in Section 4 to analyze systems where the job size has a finite support.

Theorem 2.3.2. *Consider the M/D/N queue under FCFS in the QD regime. Let Q^N denote the steady-state total number of jobs in the system. Then*

$$\mathrm{P}\{Q^N \geq N\} \sim \mathrm{P}\{Q_\infty^N \geq N\}. \tag{28}$$

Proof. The desired result (28) is equivalent to

$$\mathrm{P}\{W^N > 0\} \sim \mathrm{P}\{Q_\infty^N \geq N\}, \tag{29}$$

where W^N denotes the steady-state delay (or waiting time) in the Nth system. In what follows, we prove relation (29).

Because the steady-state delay in the M/D/N queue under FCFS has the same distribution as that in the same system under cyclic scheduling (*i.e.*, the policy under which every Nth customer is assigned to the same server; see Lemma 2 in [31]), we consider the M/D/N queue under cyclic scheduling instead, and restrict our attention to an arbitrarily chosen server in the M/D/N queue and jobs processed by this server. This subsystem is just a G/D/1 queue, where each interarrival time X_i, $i \geq 1$, is the sum of N independent copies of T_i^N's (*i.e.*, Gamma (N, λ^N))[1] and each job size is μ^{-1}. Because this G/D/1 queue has the same steady-state delay distribution as the whole M/D/N queue under cyclic scheduling, we shall analyze this subsystem and also denote its steady-state delay by W^N.

First, a fundamental result for the GI/GI/1 queue (see Proposition 1.1 of Chapter X in [2]) states that

$$W^N \stackrel{d}{=} \max_{k \geq 0} S_k, \tag{30}$$

where $\stackrel{d}{=}$ means "equal in distribution to", $S_0 = 0$ and $S_k = \sum_{i=1}^{k} (\mu^{-1} - X_i)$, $k \geq 1$.

[1] For the sake of simplicity, we suppress X_i's dependence on N by omitting the superscript N.

Therefore, we have that

$$\mathrm{P}\{S_1 > 0\} \leq \mathrm{P}\{W^N > 0\} \leq \sum_{k=1}^{\infty} \mathrm{P}\{S_k > 0\}. \tag{31}$$

Also note that, for any $k \geq 1$,

$$\begin{aligned}\mathrm{P}\{S_k > 0\} &= \mathrm{P}\left\{\sum_{i=1}^{k} X_i \leq k\mu^{-1}\right\} = \mathrm{P}\left\{\sum_{i=1}^{kN} T_i^N \leq k\mu^{-1}\right\} \\ &= \mathrm{P}\{A^N(k\mu^{-1}) \geq kN\} = \mathrm{P}\{B^{kN} \geq kN\},\end{aligned} \tag{32}$$

where B^{kN} denotes a Poisson random variable with mean $kN\rho$. The rest of the proof consists of the derivation of asymptotic lower and upper bounds that coincide in the limit.

Lower Bound: Combining $Q_\infty^N \stackrel{d}{=} B^N$, the first inequality in (31), and (32) with $k = 1$, we immediately have

$$\liminf_{N\to\infty} \frac{\mathrm{P}\{W^N > 0\}}{\mathrm{P}\{Q_\infty^N \geq N\}} \geq 1. \tag{33}$$

Upper Bound: It follows from the second inequality in (31) and (32) that

$$\begin{aligned}\mathrm{P}\{W^N > 0\} &\leq \mathrm{P}\{B^N \geq N\} + \sum_{k=2}^{\infty} \mathrm{P}\{B^{kN} \geq kN\} \\ &\leq \mathrm{P}\{B^N \geq N\} + \sum_{k=2}^{\infty} e^{-kN(\theta - \rho e^{\theta} + \rho)},\end{aligned} \tag{34}$$

for all $\theta > 0$, where the last inequality is due to the Chernoff bound. In particular, letting $\theta = -\log\rho$, which is the minimizer of $-(\theta - \rho e^{\theta} + \rho)$, yields the following upper bound

$$\mathrm{P}\{W^N > 0\} \leq \mathrm{P}\{B^N \geq N\} + \sum_{k=2}^{\infty} e^{-kNI(\rho)} \tag{35}$$

$$= \mathrm{P}\{B^N \geq N\} + \frac{e^{-2NI(\rho)}}{1 - e^{-NI(\rho)}}$$

$$\leq \mathrm{P}\{B^N \geq N\} + \frac{e^{-2NI(\rho)}}{1 - e^{-I(\rho)}}. \tag{36}$$

Denoting the constant $\left[1 - e^{-I(\rho)}\right]^{-1}$ by C_1 and dividing both sides of (36) by $\mathrm{P}\{Q_\infty^N \geq N\}$ (or equivalently $\mathrm{P}\{B^N \geq N\}$) then yields

$$\frac{\mathrm{P}\{W^N > 0)}{\mathrm{P}\{Q_\infty^N \geq N\}} \leq 1 + \frac{C_1 e^{-2NI(\rho)}}{\mathrm{P}\{Q_\infty^N \geq N\}}. \tag{37}$$

We next apply (22) to $\mathrm{P}\{Q_\infty^N \geq N\}$ on the right-hand side of (37), which gives that

$$\frac{\mathrm{P}\{W^N > 0\}}{\mathrm{P}\{Q_\infty^N \geq N\}} \leq 1 + C_1(1-\rho)\sqrt{2\pi N} \cdot e^{-NI(\rho)}[1 + o(1)], \tag{38}$$

where $a^N = o(1)$ if $\lim_{N\to\infty} a^N = 0$. Therefore,

$$\limsup_{N\to\infty} \frac{\mathrm{P}\{W^N > 0\}}{\mathrm{P}\{Q_\infty^N \geq N\}} \leq 1. \tag{39}$$

Finally, combining (33) and (39) completes the proof. □

2.3.3 Random walk estimates and bounds for the M/D/c queue

In this subsection, we provide a random walk result and two bound results for the M/D/c queue which will be relied upon in the proof of our main results. Because we shall apply these results to subsystems with different number of servers (*i.e.*, $s_i(N)$, $i = 1, ..., m(N)-1$), we denote the number of servers by c, instead of N, in stating them.

First, we state a result concerning the maximum of a random walk process. Its proof is deferred to the end of this subsection.

Proposition 2.3.3. *Let $\{B_i, i \in \mathbb{N}\}$ be a sequence of i.i.d. Poisson random variables with mean ρc, where $c \in \mathbb{N}$ and $\rho \in (0,1)$. Let $M = \max_{n\geq 0} S_n$, with $S_0 := 0$ and $S_n := \sum_{i=1}^n (B_i - c)$ for $n \geq 1$.* [1] *Then for any $c \in \mathbb{N}$,*

$$\mathrm{P}\{M \geq j\} \leq K(\rho)\rho^j \cdot \mathrm{P}\{M \geq 1\}, \quad \textit{for all } j \in \mathbb{N}, \tag{40}$$

where $K(\rho)$ is a function of ρ only and, in particular, $K(\rho)$ is independent of c.

Now, utilizing Proposition 2.3.3, we prove some bounds on the steady-state distribution of the total number of jobs in an M/D/c queue.

Lemma 2.3.4. *Consider an M/D/c queue with the traffic intensity $\rho \in (0,1)$. Let Q denote the steady-state total number of jobs in the system. Then for all $j \in \mathbb{N}$,*

$$\mathrm{P}\{Q = c - j\} \leq \rho^{-j} \cdot \mathrm{P}\{Q = c\}, \tag{41}$$

$$\mathrm{P}\{Q \geq c + j\} \leq K(\rho)\rho^j \cdot \mathrm{P}\{Q \geq c\}, \tag{42}$$

where $K(\rho)$ is a function of ρ only.

[1] *We suppress the dependence of B_i's, S_n's and M on c in the notation.*

Proof. In the M/D/c queue, the following relation holds (see [51])

$$Q \stackrel{d}{=} (Q-c)^+ + B, \tag{43}$$

where B is Poisson distributed with mean ρc. Specifically, for any $n \geq 0$,

$$\mathrm{P}\{Q=n\} = \frac{e^{-\rho c}(\rho c)^n}{n!} \cdot \sum_{k=0}^{c} \mathrm{P}\{Q=k\} + \sum_{k=c+1}^{c+n} \mathrm{P}\{Q=k\} \cdot \frac{e^{-\rho c}(\rho c)^{n-k+c}}{(n-k+c)!}. \tag{44}$$

We first verify (41). Substituting n in (44) with c and $c-j$ respectively yields that

$$\mathrm{P}\{Q=c\} = \frac{e^{-\rho c}(\rho c)^c}{c!} \cdot \sum_{k=0}^{c} \mathrm{P}\{Q=k\} + \sum_{k=c+1}^{2c} \mathrm{P}\{Q=k\} \cdot \frac{e^{-\rho c}(\rho c)^{2c-k}}{(2c-k)!}, \tag{45}$$

and for all $j = 1, 2, ..., c$,

$$\mathrm{P}\{Q=c-j\} = \frac{e^{-\rho c}(\rho c)^{c-j}}{(c-j)!} \cdot \sum_{k=0}^{c} \mathrm{P}\{Q=k\} + \sum_{k=c+1}^{2c-j} \mathrm{P}\{Q=k\} \cdot \frac{e^{-\rho c}(\rho c)^{2c-j-k}}{(2c-j-k)!}. \tag{46}$$

Now define

$$A := \frac{e^{-\rho c}(\rho c)^c}{c!} \cdot \sum_{k=0}^{c} \mathrm{P}\{Q=k\}, \quad B_k := \mathrm{P}\{Q=k\} \cdot \frac{e^{-\rho c}(\rho c)^{2c-k}}{(2c-k)!}, \text{ for all } k = c+1, ..., 2c, \tag{47}$$

$$C := \frac{e^{-\rho c}(\rho c)^{c-j}}{(c-j)!} \cdot \sum_{k=0}^{c} \mathrm{P}\{Q=k\}, \quad D_k := \mathrm{P}\{Q=k\} \cdot \frac{e^{-\rho c}(\rho c)^{2c-j-k}}{(2c-j-k)!}. \tag{48}$$

Substituting these definitions into (45) and (46) respectively then yields

$$\mathrm{P}\{Q=c\} = A + \sum_{k=c+1}^{2c} B_k \geq A + \sum_{k=c+1}^{2c-j} B_k, \text{ for all } j = 1, 2, ..., c, \tag{49}$$

and

$$\mathrm{P}\{Q=c-j\} = C + \sum_{k=c+1}^{2c-j} D_k. \tag{50}$$

Because $A/C \geq \rho^j$ and $B_k/D_k \geq \rho^j$ for all $k = c+1, ..., 2c-j$, (41) follows from dividing (49) by (50).

Next, we turn to proving (42). From (43), we have

$$(Q-c)^+ \stackrel{d}{=} [(Q-c)^+ + B - c]^+, \tag{51}$$

or

$$Q_q \stackrel{d}{=} (Q_q + B - c)^+, \tag{52}$$

where Q_q denotes the steady-state number of jobs waiting in the queue. We further let $M = \max_{n\geq 0} S_n$, where $S_0 = 0$, $S_n = \sum_{i=1}^{n}(B_i - c)$ for $n \geq 1$, and B_i's are independent random variables equal in distribution to B. By the standard Lindley recursion result (see Corollary 6.6 on page 94 of [2]), we have $Q_q \stackrel{d}{=} M$. Then we apply Proposition 2.3.3 to Q_q and arrive at

$$\mathrm{P}\{Q_q \geq j\} \leq K(\rho)\rho^j \cdot \mathrm{P}\{Q_q \geq 1\}, \quad \text{for all } j \in \mathbb{N},$$

or equivalently,

$$\mathrm{P}\{Q \geq c + j\} \leq K(\rho)\rho^j \cdot \mathrm{P}\{Q_q \geq 1\}, \quad \text{for all } j \in \mathbb{N}. \tag{53}$$

Finally, (42) follows from (53) because $\mathrm{P}\{Q_q \geq 1\} \leq \mathrm{P}\{Q \geq c\}$. □

Our next preparatory result is an exponential upper bound on $\mathrm{P}\{Q \geq c\}$ in the M/D/c queue. This bound is not used in the proof for systems where the job size has a finite support, but is needed in our proof for general (discrete) job size distributions.

Lemma 2.3.5. *Consider an M/D/c queue with the traffic intensity $\rho \in (0, 1)$. Let Q denote the steady-state total number of jobs in the system. Then*

$$\mathrm{P}\{Q \geq c\} \leq \frac{e^{-cI(\rho)}}{1 - e^{-I(\rho)}}. \tag{54}$$

Proof. The proof is similar to the derivation of (35) from (34). Here, in addition to applying the minimizing Chernoff bound to the summation from $k = 2$ to ∞, we apply it to the first term of (34) as well. Therefore, with N in (35) replaced by c, we obtain that

$$\mathrm{P}\{Q \geq c\} \leq \sum_{k=1}^{\infty} e^{-kcI(\rho)} = \frac{e^{-cI(\rho)}}{1 - e^{-cI(\rho)}} \leq \frac{e^{-cI(\rho)}}{1 - e^{-I(\rho)}}. \tag{55}$$

□

Finally, we prove the random walk estimate result. Our proof is related to the light traffic analysis of random walks (see [2]).

Proof of Proposition 2.3.3. Step 1. We prove an inequality concerning the increment of the random walk (*i.e.*, $B_1 - c$): for any $c \in \mathbb{N}$,

$$\mathrm{P}\{B_1 - c \geq j\} \leq K_1(\rho)\rho^j \cdot \mathrm{P}\{B_1 - c \geq 1\}, \quad \text{for all } j \in \mathbb{N}, \tag{56}$$

where $K_1(\rho)$ is a function of ρ only.

From (25) (with Q_∞^N and N replaced by B_1 and c respectively), we obtain that

$$\mathrm{P}\{B_1 - c \geq j\} \leq \frac{\rho^j}{1-\rho} \cdot \mathrm{P}\{B_1 = c\}, \quad \text{for all } j \in \mathbb{N}, \tag{57}$$

and

$$\mathrm{P}\{B_1 - c \geq 1\} \geq \sum_{i=1}^{\infty} \left(\frac{c\rho}{c+i}\right)^i \cdot \mathrm{P}\{B_1 = c\} \geq \sum_{i=1}^{\infty} \left(\frac{\rho}{1+i}\right)^i \cdot \mathrm{P}\{B_1 = c\} \geq \frac{\rho}{2} \cdot \mathrm{P}\{B_1 = c\}. \tag{58}$$

Then, dividing (57) by (58) yields (56), where $K_1(\rho) = 2[(1-\rho)\rho]^{-1}$.

Step 2. We prove a similar bound concerning the first ascending ladder height of the random walk, namely, S_{τ_+}, where $\tau_+ = \inf\{n \geq 1 : S_n > 0\}$. Specifically, we show that, for any $c \in \mathbb{N}$,

$$\mathrm{P}\{S_{\tau_+} \geq j\} \leq K_2(\rho)\rho^j \cdot \mathrm{P}\{S_{\tau_+} \geq 1\}, \quad \text{for all } j \in \mathbb{N}, \tag{59}$$

where $K_2(\rho)$ is a function of ρ only. Note that S_{τ_+} is defined as 0 when $\tau_+ = \infty$.

Following the same notation as used on pages 221 and 223 of [2], we define $\tau_- := \inf\{n \geq 1 : S_n \leq 0\}$ and $\tau_-(i+1) := \inf\{n > \tau_-(i) : S_n \leq S_{\tau_-(i)}\}$, where $\tau_-(1) := \tau_-$. Note that, the descending ladder heights are not strict; in particular, unlike S_{τ_+}, S_{τ_-} can be 0, even when $\tau_- < \infty$.

First, we apply equation (1.7) on page 269 of [2] [2], with the set A in that expression replaced by $[j, \infty)$, and obtain that

$$\begin{aligned}
\mathrm{P}\{S_{\tau_+} \geq j\} &= \sum_{x=-\infty}^{0} \mathrm{P}\{B_1 - c \geq j - x\} \cdot \left[\mathbf{1}\{x = 0\} + \sum_{i=1}^{\infty} \mathrm{P}\{S_{\tau_-(i)} = x\}\right] \\
&= \mathrm{P}\{B_1 - c \geq j\} + \sum_{x=-\infty}^{0} \mathrm{P}\{B_1 - c \geq j - x\} \cdot \sum_{i=1}^{\infty} \mathrm{P}\{S_{\tau_-(i)} = x\} \\
&= \mathrm{P}\{B_1 - c \geq j\} + \sum_{k=0}^{\infty} R(k) \cdot \mathrm{P}\{B_1 - c \geq j + k\},
\end{aligned} \tag{60}$$

[2] Note that in their notation G_+ corresponds to the distribution of S_{τ_+}, *i.e.*, for set A, $G_+(A) = \mathrm{P}\{S_{\tau_+} \in A\}$. Also, for set A, $U_-(A) := \sum_{i=0}^{\infty} G_-^{*i}(A)$, where $G_-^{*0}(A) = \mathbf{1}\{0 \in A\}$ and G_-^{*i}, $i \geq 1$, is the ith convolution of G_-, with G_- being the distribution of S_{τ_-}; more specifically, $U_-(A) = \mathbf{1}\{0 \in A\} + \sum_{i=1}^{\infty} \mathrm{P}\{S_{\tau_-(i)} \in A\}$, by noting that for all $i \geq 1$, G_-^{*i} is exactly the distribution of $S_{\tau_-(i)}$. Also F in their expression (1.7) denotes the distribution of the random walk increment, which means in our case $F(A) = \mathrm{P}\{B_1 - c \in A\}$.

where $R(k) := \sum_{i=1}^{\infty} \mathrm{P}\{S_{\tau_-(i)} = -k\}$, for all $k \in \mathbb{Z}^+$.

Next, we establish an upper bound of $R(k)$, which is uniform in $k \in \mathbb{Z}^+$. For all $k \in \mathbb{Z}^+$, we have

$$\begin{aligned} R(k) &= E[\#\{i \geq 1 : S_{\tau_-(i)} = -k\}] \\ &\leq E[\#\{i \geq 1 : S_{\tau_-(i)} = -k\} | S_{\tau_-(i)} = -k \text{ for some } i \in \mathbb{N}] \\ &= 1 + \sum_{i=1}^{\infty} i \cdot \mathrm{P}\{S_{\tau_-} = 0\}^i \cdot \mathrm{P}\{S_{\tau_-} < 0\} \qquad (61) \\ &\leq 1 + \sum_{i=1}^{\infty} i \cdot \mathrm{P}\{S_{\tau_-} = 0\}^i, \qquad (62) \end{aligned}$$

where (61) holds because $\{S_{\tau_-(i+1)} - S_{\tau_-(i)}, i \geq 1\}$ is a sequence of i.i.d. random variables equal in distribution to S_{τ_-}.

Also, since $\{S_{\tau_-} = 0\} \subset \{B_1 \geq c\}$, we have that

$$\mathrm{P}\{S_{\tau_-} = 0\} \leq \mathrm{P}\{B_1 \geq c\} \leq e^{-cI(\rho)} \leq e^{-I(\rho)} < 1, \qquad (63)$$

where the third last inequality follows from applying the minimizing Chernoff bound (in the same way as we obtain (35)). Then with $\delta_1(\rho) := e^{-I(\rho)}$, combining (62) and (63) leads to the following uniform bound

$$\begin{aligned} R(k) &\leq 1 + \sum_{i=1}^{\infty} i \cdot \delta_1(\rho)^i \\ &= \delta_2(\rho), \quad \text{for all } k \in \mathbb{Z}^+, \qquad (64) \end{aligned}$$

where $\delta_2(\rho) = 1 + \delta_1(\rho) \cdot [1 - \delta_1(\rho)]^{-2} > 1$.

Finally, combining (56), (60), (64) and using $\mathrm{P}\{B_1 - c \geq 1\} \leq \mathrm{P}\{S_{\tau_+} \geq 1\}$ yields (59) as follows

$$\begin{aligned} \mathrm{P}\{S_{\tau_+} \geq j\} &\leq K_1(\rho)\rho^j \cdot \mathrm{P}\{B_1 - c \geq 1\} + \sum_{k=0}^{\infty} \delta_2(\rho) \cdot K_1(\rho)\rho^{j+k} \cdot \mathrm{P}\{B_1 - c \geq 1\} \\ &\leq K_1(\rho)\rho^j \cdot \mathrm{P}\{S_{\tau_+} \geq 1\} + \sum_{k=0}^{\infty} \delta_2(\rho) \cdot K_1(\rho)\rho^{j+k} \cdot \mathrm{P}\{S_{\tau_+} \geq 1\} \\ &= K_2(\rho)\rho^j \cdot \mathrm{P}\{S_{\tau_+} \geq 1\}, \end{aligned}$$

where $K_2(\rho) = K_1(\rho) + \delta_2(\rho)K_1(\rho) \cdot (1 - \rho)^{-1}$.

Step 3. We eventually prove (40). First, there exists a constant $\beta \in (0, \rho)$, whose value is independent of c, such that

$$\mathrm{E}\Big[\beta^{-(B_1-c)}\Big] = 1. \tag{65}$$

Specifically, using the fact that B_1 is Poisson with mean ρc, a straightforward calculation shows that the desired β satisfies $\rho(1 - 1/\beta) = \log \beta$.

Next, using $\beta \in (0, \rho)$ and applying the Kolmogorov's Inequality for (Sub)Martingales to $\{\beta^{-S_n}, n \in \mathbb{Z}_+\}$, we have

$$\mathrm{P}\{M \geq j\} = \mathrm{P}\{\beta^{-M} \geq \beta^{-j}\} \leq \beta^j, \quad \text{for all } j \in \mathbb{Z}_+. \tag{66}$$

In other words, $M \leq_{\mathrm{st}} M_\beta$, where the subscript $_{\mathrm{st}}$ stands for the usual stochastic order and $\mathrm{P}\{M_\beta = j\} = (1 - \beta)\beta^j$, for all $j \in \mathbb{Z}^+$.

Finally, using $\{S_{\tau_+} \geq 1\} = \{M \geq 1\}$ and $M|S_{\tau_+} \geq 1 \stackrel{d}{=} G + M$, with $G := S_{\tau_+}|S_{\tau_+} \geq 1$, we obtain that, for all $j \in \mathbb{N}$,

$$\begin{aligned}
\mathrm{P}\{M \geq j | M \geq 1\} &= \mathrm{P}\{G + M \geq j\} \leq \mathrm{P}\{G + M_\beta \geq j\} \\
&= \sum_{k=0}^{j-1} \mathrm{P}\{M_\beta = k\} \cdot \mathrm{P}\{G \geq j - k\} + \mathrm{P}\{M_\beta \geq j\} \\
&= \sum_{k=0}^{j-1} \mathrm{P}\{M_\beta = k\} \cdot \mathrm{P}\{S_{\tau_+} \geq j - k | S_{\tau_+} \geq 1\} + \mathrm{P}\{M_\beta \geq j\} \\
&\leq \sum_{k=0}^{j-1} (1 - \beta)\beta^k \cdot K_2(\rho)\rho^{j-k} + \beta^j \qquad (67) \\
&< \rho^j \sum_{k=0}^{\infty} (1 - \beta) \left(\frac{\beta}{\rho}\right)^k K_2(\rho) + \rho^j \\
&= K(\rho)\rho^j, \qquad (68)
\end{aligned}$$

where (67) follows from (59) and $K(\rho) = K_2(\rho)(1-\beta) \cdot (1 - \beta\rho^{-1})^{-1} + 1$. This is equivalent to (40). □

2.3.4 Combinatorial results

This subsection contains two combinatorial results, which are used in the proof of our main theorems. We start by giving some necessary notation. Throughout this chapter, we use

the overline (or bar) symbol to denote vectors, *e.g.*, $\overline{c} = (c_1, c_2, ..., c_{m-1}, c_m)$. For any $\overline{c} = (c_1, c_2, ..., c_{m-1}, c_m) \in \mathbb{Z}_+^m$, we define $\| \overline{c} \| := \sum_{i=1}^m c_i$. For any $\overline{c} \in \mathbb{Z}_+^m$ and $j \in \mathbb{Z}_+$,

$$S(\overline{c}, j) := \{\overline{x} \in \mathbb{Z}_+^m : \| \overline{x} \| = \| \overline{c} \| + j\}, \tag{69}$$

or in words, $S(\overline{c}, j)$ denotes the set of nonnegative integer solutions to $\| \overline{x} \| = \| \overline{c} \| + j$. The first combinatorial result that we use in later proofs is on the cardinality of $S(\overline{c}, j)$. From page 15 of [49], we have that

$$|S(\overline{c}, j)| = \binom{\| \overline{c} \| + j + m - 1}{m - 1}. \tag{70}$$

The second combinatorial result that we need (see page 1077 of [22]) is

$$\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k, \quad \text{for all } k = 0, ..., n. \tag{71}$$

2.4 Job sizes with finite support

In this section, we consider the case in which the job size S can take on finitely many possible values and prove Theorem 2.2.3. The following lemma is used in our proof of Theorem 2.2.3. We include its proof in Section 2.8.

Lemma 2.4.1. *Consider m independent $M/D/c_i$ queues, $i = 1, ..., m$, all with the same traffic intensity ρ_o. Let Q_i denote the steady-state total number of jobs in the $M/D/c_i$ queue, $i = 1, ..., m$. Then for all $j \in \mathbb{Z}_+$,*

$$\max_{\overline{n} \in S(\overline{c}, j)} \mathrm{P}\{Q_i = n_i, i = 1, ..., m\} \leq K(\rho_o)^m \rho_o^j \cdot \mathrm{P}\{Q_i \geq c_i, \textit{for all } i = 1, ..., m\}, \tag{72}$$

where $K(\cdot)$ is the same as given in Lemma 2.3.4.

Next, we prove Theorem 2.2.3.

Proof of Theorem 2.2.3. Let Q^N be the steady-state total number of jobs in the Nth system under policy $\pi(N)$. We need to prove (1). Due to (4) and (23), it suffices to show that

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathrm{P}\{Q^N \geq N\} \leq -I(\rho). \tag{73}$$

First, from (8) and (9) we easily have that $\lim_{N\to\infty}\rho_i(N)=\rho$, for $i=1,...,m-1$, and $\rho_m(N)\le\rho$. Then for any $\epsilon\in(0,1-\rho)$, there exists an $N_\epsilon\in\mathbb{N}$ such that, for any $N>N_\epsilon$,

$$\rho_i(N)\le\rho_\epsilon,\quad \text{for all } i=1,...,m, \tag{74}$$

where $\rho_\epsilon:=\rho(1-\epsilon)^{-1}$.

Let us fix an $\epsilon\in(0,1-\rho)$ and construct an auxiliary sequence, for which we only define the Nth system if $N\ge N_\epsilon$. In the Nth system of the auxiliary sequence, there are N servers and jobs arrive according to the process $\{A^N(t),t\ge 0\}$. A job is type i with probability p_i, and type i jobs have size $s_i(N)\rho_\epsilon\cdot(\lambda^N p_i)^{-1}$, $i=1,...,m$. Also $s_i(N)$ of the N servers are dedicated to processing type i jobs, $i=1,...,m$, on an FCFS basis. In other words, the Nth system of the auxiliary sequence is the same the original Nth system, except that the size of type i jobs is $s_i(N)\rho_\epsilon\cdot(\lambda^N p_i)^{-1}$, instead of d_i, $i=1,...,m$.

By basic properties of the Poisson process, for systems either in the original sequence or the auxiliary sequence, the job arrivals to the m server pools are independent Poisson processes with rate $\lambda^N p_i$, $i=1,...,m$. Therefore, the m server pools can be viewed as m subsystems operating completely independently. Furthermore, each subsystem is an M/D/$s_i(N)$ queue.

By the construction of the auxiliary sequence, all the m server pools in each auxiliary system have the same traffic intensity ρ_ϵ. Since all subsystems are M/D/$s_i(N)$ queues and each subsystem in the Nth auxiliary system has the same number of servers and job arrival process as the corresponding one in the Nth original system, (74) then implies

$$Q_i^N\le_{\text{st}} Q_{i,u}^N,\quad \text{for all } i=1,...,m, \tag{75}$$

where Q_i^N denotes the steady-state total number of type i jobs in the Nth original system and $Q_{i,u}^N$ denotes the steady-state total number of type i jobs in the Nth auxiliary system. Therefore,

$$Q^N=\sum_{i=1}^m Q_i^N\le_{\text{st}}\sum_{i=1}^m Q_{i,u}^N. \tag{76}$$

Define $\overline{s_N}:=(s_1(N),s_2(N),...,s_{m-1}(N),s_m(N))$. For any $N>N_\epsilon$, it follows from (76)

that

$$\begin{aligned}\mathrm{P}\{Q^N \geq N\} = \mathrm{P}\left\{\sum_{i=1}^{m} Q_i^N \geq N\right\} &\leq \mathrm{P}\left\{\sum_{i=1}^{m} Q_{i,u}^N \geq N\right\} \\ &= \sum_{j=0}^{\infty} \mathrm{P}\left\{\sum_{i=1}^{m} Q_{i,u}^N = N+j\right\} \\ &= \sum_{j=0}^{\infty} \sum_{\overline{n}\in S(\overline{s_N},j)} \mathrm{P}\left\{Q_{i,u}^N = n_i, i=1,...,m\right\} \qquad (77)\end{aligned}$$

From (70), we know

$$|S(\overline{s_N}, j)| = \binom{N+j+m-1}{m-1}. \qquad (78)$$

Applying Lemma 2.4.1 and (78) to (77) then leads to

$$\mathrm{P}\{Q^N \geq N\} \leq \sum_{j=0}^{\infty} \binom{N+j+m-1}{m-1} K(\rho_\epsilon)^m \rho_\epsilon^j \cdot \mathrm{P}\{Q_{i,u}^N \geq s_i(N), \text{for all } i=1,...,m\}. \quad (79)$$

We next apply (71) to (79) and obtain that

$$\begin{aligned}\mathrm{P}\{Q^N \geq N\} &\leq \sum_{j=0}^{\infty} \left[\frac{(N+j+m-1)e}{m-1}\right]^{m-1} K(\rho_\epsilon)^m \rho_\epsilon^j \cdot \mathrm{P}\{Q_{i,u}^N \geq s_i(N), \text{for all } i=1,...,m\} \\ &= \frac{K(\rho_\epsilon)^m e^{m-1} \cdot \mathrm{P}\{Q_{i,u}^N \geq s_i(N), \text{for all } i=1,...,m\} \cdot N^{m-1}}{(m-1)^{m-1}(1-\rho_\epsilon)}[1+o(1)], \quad (80)\end{aligned}$$

where the last equality follows from $\sum_{j=0}^{\infty}(N+j+m-1)^{m-1}\rho_\epsilon^j = \frac{N^{m-1}}{1-\rho_\epsilon}[1+o(1)]$ by the Monotone Convergence Theorem.

From (22), (28), $s_i(N) \sim Np_id_i\mu$, and the independence among subsystems, we have

$$\begin{aligned}\mathrm{P}\{Q_{i,u}^N \geq s_i(N), \text{for all } i=1,...,m\} &= \prod_{i=1}^{m} \mathrm{P}\{Q_{i,u}^N \geq s_i(N)\} \\ &= \frac{1}{(2\pi N)^{\frac{m}{2}}(1-\rho_\epsilon)^m \sqrt{\prod_{i=1}^{m} p_i d_i \mu}} \cdot e^{-NI(\rho_\epsilon)}[1+o(1)]. \qquad (81)\end{aligned}$$

Combining (80) and (81) then yields that

$$\mathrm{P}\{Q^N \geq N\} \leq \frac{K(\rho_\epsilon)^m e^{m-1} N^{\frac{m}{2}-1}}{(2\pi)^{\frac{m}{2}}(m-1)^{m-1}(1-\rho_\epsilon)^{m+1}\sqrt{\prod_{i=1}^{m} p_i d_i \mu}} \cdot e^{-NI(\rho_\epsilon)}[1+o(1)]. \qquad (82)$$

Taking logarithms, dividing both sides of (82) by N, and letting $N \to \infty$, we have

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q^N \geq N\} \leq -I(\rho_\epsilon). \qquad (83)$$

Finally, letting $\epsilon \to 0$ in (83) yields (73) and this completes the proof. □

From Theorem 2.2.3 and its proof, it is easy to deduce that under policy $\pi(N)$,

$$\lim_{N\to\infty}\frac{1}{N}\log \mathrm{P}\{Q^N \geq N\} = \lim_{N\to\infty}\frac{1}{N}\log \mathrm{P}\{Q_i^N \geq s_i(N), \text{for all } i = 1, ..., m\}. \tag{84}$$

This suggests that, when the system operates under the load-balancing SITA, the most likely way for the event $\{Q^N \geq N\}$ to happen is the occurrence of $\{Q_i^N \geq s_i(N), \text{for all } i = 1, ..., m\}$.

Remark 2.4.2. For systems with m classes of customers, where class i customers' service times are exponential with mean d_i, $i = 1, ..., m$, a similar load-balancing size-based (or equivalently class-based) task assignment policy is strongly optimal in the QD regime. This type of model and its variants, which are often proper for service systems with human customers, have been studied in other contexts; for example, see [50] and the references therein.

Specifically, if a customer belongs to class i with probability p_i, $i = 1, ..., m$, the optimal policy is just to allocate $s_i(N)$ servers for serving class i customers exclusively, where $s_i(N)$'s are the same as defined in (7). The strong optimality holds in this case, because most results that we have proved so far for M/D/$\cdot$ queues also hold true if the service times or job sizes are exponentials.

More specifically, for the M/M/N queue, (28) can be proved as follows. With $B^N \sim$ Poisson$(N\rho)$, we rewrite expression (2.2) in [56] as

$$\begin{aligned}\mathrm{P}\{Q^N \geq N\} &= \frac{\mathrm{P}\{B^N = N\}}{\rho\mathrm{P}\{B^N = N\} + (1-\rho)\mathrm{P}\{B^N \leq N\}} \\ &= \frac{\mathrm{P}\{B^N = N\}}{o(1) + (1-\rho)[1+o(1)]} \\ &= \frac{\mathrm{P}\{B^N = N\}}{1-\rho}[1+o(1)],\end{aligned} \tag{85}$$

where the second last equality holds, because $\mathrm{P}\{B^N = N\} = o(1)$ and $\mathrm{P}\{B^N \leq N\} = 1 + o(1)$ by applying the Weak Law of Large Numbers to B^N (which is equal to the sum of N independent Poisson(ρ) random variables). Then combining (27) and (85) yields (28). For the M/M/c queue, both (41) and (42) easily follow from the exact formula for the distribution of Q. In fact, (42) holds as an equality with $K(\rho) = 1$, in that case. As a

consequence, Lemma 2.4.1 also holds for m M/M/$\cdot$ queues with $K(\rho_0) = 1$; indeed, one may prove a stronger result:

$$\max_{\overline{n} \in S(\overline{c},j)} \mathrm{P}\{Q_i = n_i, i = 1, ..., m\} = \rho_o^j \cdot \mathrm{P}\{Q_i = c_i, \text{for all } i = 1, ..., m\}. \tag{86}$$

Therefore, the whole proof of Theorem 2.2.3 can be applied with minor changes to establish the strong optimality of the load-balancing size-based (or class-based) task assignment policy.

2.5 Discrete job sizes

In this section, we prove Theorem 2.2.4, which implies Corollary 2.2.5. We start by giving an intuitive explanation of the policy construction.

With respect to the performance measure P{steady-state total number of jobs in the system $\geq$ total number of servers}, we learn from Theorem 2.2.3 that finitely many load-balanced M/D/$\cdot$ systems perform as well as an infinite-server queue in the QD regime. In fact, in the proof of Theorem 2.2.3, the step from (82) to (83) would still hold if m were replaced by some other $o(N/\log N)$ quantity, and thus $o(N/\log N)$ many load-balanced M/D/$\cdot$ systems also perform as well as an infinite-server queue. This observation leads to our choice of $m(N)$ (see (15)) at the order of N^η for some $\eta \in (0, 1)$, where $N^\eta = o(N/\log N)$.

Therefore, the main idea is to construct a family of SITA policies, under which the original system consists of two components: the first component can be bounded from above (with respect to our performance measure) by $o(N/\log N)$ many load-balanced M/D/$\cdot$ queues, and the second component is a single-server queue, where this one server only processes jobs with extremely large sizes — specifically, growing super-exponentially fast as $N \uparrow \infty$, cf. (110) — such that the contribution from the single-server queue to the total system size is negligible (cf. (104)). Also, note that the assumption on the finiteness of the αth moment of S, for $\alpha > 1$, provides us with a control on the tail distribution of the job size, which is critically relied upon in the policy construction and the proof.

Before analyzing the performance of the policy $\pi_\epsilon(N)$, we first show that it is feasible for a large enough N.

Proposition 2.5.1. *For any $\epsilon > 0$, there exists $N_{\epsilon,1} \in \mathbb{N}$, such that, for all $N > N_{\epsilon,1}$, the policy $\pi_\epsilon(N)$ prescribed by Algorithm 1 is a feasible SITA policy.*

Proof. First, by the continuity of $I(\cdot)$ and its monotonicity in the interval $(0,1)$, for any $\epsilon > 0$, there exists $\sigma(\epsilon) \in (0, 1-\rho)$, such that, with $\rho_\epsilon := \rho[1-\sigma(\epsilon)]^{-1}$, (12) holds.

Next, we show that $\{r_i(N), i = 1, ..., 2\lceil N^\eta \rceil\}$ is increasing in i. It is sufficient to show $\{f_i(N), i = 1, ..., 2\lceil N^\eta \rceil - 1\}$ is increasing in i. It easily follows from (13) that

$$f_1(N) < ... < f_{\lceil N^\eta \rceil}(N). \tag{87}$$

From the recursive definition (14), for $i = \lceil N^\eta \rceil, ..., 2\lceil N^\eta \rceil - 1$,

$$f_i(N) = \lceil N^\eta \rceil^{\alpha^{i-\lceil N^\eta \rceil}} N^{-\gamma[\alpha^{i-\lceil N^\eta \rceil - 1} + 2\alpha^{i-\lceil N^\eta \rceil - 2} + \cdots + (i-\lceil N^\eta \rceil - 1)\alpha + (i-\lceil N^\eta \rceil)]}. \tag{88}$$

From (14) and (88), we have, for all $i = \lceil N^\eta \rceil + 1, ..., 2\lceil N^\eta \rceil - 1$,

$$\begin{aligned}
\frac{f_i(N)}{f_{i-1}(N)} &= f_{i-1}^{\alpha-1}(N) N^{-(i-\lceil N^\eta \rceil)\gamma} \\
&\geq \frac{\left(N^{\eta\alpha^{i-\lceil N^\eta \rceil - 1} - \gamma[\alpha^{i-\lceil N^\eta \rceil - 2} + 2\alpha^{i-\lceil N^\eta \rceil - 3} + \cdots + (i-\lceil N^\eta \rceil - 2)\alpha + (i-\lceil N^\eta \rceil - 1)]}\right)^{\alpha-1}}{N^{(i-\lceil N^\eta \rceil)\gamma}} \\
&= N^{\eta\alpha^{i-\lceil N^\eta \rceil - 1}(\alpha-1) - \gamma\frac{\alpha^{i-\lceil N^\eta \rceil}-1}{\alpha-1}}.
\end{aligned} \tag{89}$$

Because (11) implies that

$$\gamma < \frac{\eta(\alpha-1)^2}{\alpha} \cdot \frac{\alpha^{i-\lceil N^\eta \rceil}}{\alpha^{i-\lceil N^\eta \rceil} - 1}, \quad \text{for all } i = \lceil N^\eta \rceil + 1, ..., 2\lceil N^\eta \rceil - 1, \tag{90}$$

the exponent in (89) is positive, namely

$$\eta\alpha^{i-\lceil N^\eta \rceil - 1}(\alpha-1) - \gamma\frac{\alpha^{i-\lceil N^\eta \rceil}-1}{\alpha-1} > 0 \tag{91}$$

and thus $f_i(N)/f_{i-1}(N) > 1$, or $f_i(N)$ is increasing in i, for all $i = \lceil N^\eta \rceil, ..., 2\lceil N^\eta \rceil - 1$. This, together with (87), establishes the assertion that $\{f_i(N), i = 1, ..., 2\lceil N^\eta \rceil - 1\}$ is increasing in i.

To check feasibility, the last condition that we need to verify is that $s_{2\lceil N^\eta \rceil - 1}(N)$ as defined by (18) is positive. We shall show that, for any $\epsilon > 0$, there exists $N_{\epsilon,1} \in \mathbb{N}$, such that, for all $N > N_{\epsilon,1}$,

$$s_{2\lceil N^\eta \rceil - 1}(N) - \lceil N P_{L_{2\lceil N^\eta \rceil - 1}} f_{2\lceil N^\eta \rceil - 1}(N) \mu(1-\sigma(\epsilon)) \rceil > 0, \tag{92}$$

or

$$N-1-\sum_{i=1}^{2\lceil N^\eta\rceil-2} s_i(N) - \lceil NP_{L_{2\lceil N^\eta\rceil-1}} f_{2\lceil N^\eta\rceil-1}(N)\mu(1-\sigma(\epsilon))\rceil > 0. \tag{93}$$

Inequality (92) is obviously a stronger statement than $s_{2\lceil N^\eta\rceil-1}(N) > 0$ and we shall need it in later proofs.

Define

$$\Sigma_0 := \sum_{i=1}^{2\lceil N^\eta\rceil-2} s_i(N) + \lceil NP_{L_{2\lceil N^\eta\rceil-1}} f_{2\lceil N^\eta\rceil-1}(N)\mu(1-\sigma(\epsilon))\rceil. \tag{94}$$

From (17) and (19), we have that

$$\begin{aligned}\Sigma_0 &= \sum_{i=1}^{\lceil N^\eta\rceil} s_i(N) + \left(\sum_{i=\lceil N^\eta\rceil+1}^{2\lceil N^\eta\rceil-2} s_i(N) + \lceil NP_{L_{2\lceil N^\eta\rceil-1}} f_{2\lceil N^\eta\rceil-1}(N)\mu(1-\sigma(\epsilon))\rceil\right)\\ &\le \Big(N(1-\sigma(\epsilon)) + \lceil N^\eta\rceil\Big) + \left(N\mu(1-\sigma(\epsilon))\sum_{i=\lceil N^\eta\rceil+1}^{2\lceil N^\eta\rceil-1} P_{L_i} f_i(N) + \lceil N^\eta\rceil - 1\right) \qquad (95)\\ &= N(1-\sigma(\epsilon)) + 2\lceil N^\eta\rceil - 1 + N\mu(1-\sigma(\epsilon))\sum_{i=\lceil N^\eta\rceil+1}^{2\lceil N^\eta\rceil-1} P_{L_i} f_i(N), \qquad (96)\end{aligned}$$

where the terms $\lceil N^\eta\rceil$ and $\lceil N^\eta\rceil - 1$ in (95) count the maximum possible rounding error. Furthermore,

$$\sum_{i=\lceil N^\eta\rceil+1}^{2\lceil N^\eta\rceil-1} P_{L_i} f_i(N) \le \mathrm{E}[S^\alpha]\cdot \sum_{i=\lceil N^\eta\rceil+1}^{2\lceil N^\eta\rceil-1} f_{i-1}^{-\alpha}(N) f_i(N), \tag{97}$$

because by the Markov inequality,

$$P_{L_i} = \mathrm{P}\{S\in(f_{i-1}(N), f_i(N)]\} < \mathrm{P}\{S > f_{i-1}(N)\} \le \mathrm{E}[S^\alpha]\cdot f_{i-1}^{-\alpha}(N). \tag{98}$$

Also, from (14) we have

$$\sum_{i=\lceil N^\eta\rceil+1}^{2\lceil N^\eta\rceil-1} f_{i-1}^{-\alpha}(N) f_i(N) = \sum_{i=\lceil N^\eta\rceil+1}^{2\lceil N^\eta\rceil-1} N^{-(i-\lceil N^\eta\rceil)\gamma} \le \sum_{i=1}^{\infty} N^{-i\gamma}. \tag{99}$$

By applying (97) and (99) to (96), we obtain that

$$\begin{aligned}\Sigma_0 &\le N(1-\sigma(\epsilon)) + 2\lceil N^\eta\rceil - 1 + N\mu(1-\sigma(\epsilon))\mathrm{E}[S^\alpha]\cdot\sum_{i=1}^{\infty} N^{-i\gamma}\\ &= N(1-\sigma(\epsilon)) + o(N) < N-1, \qquad (100)\end{aligned}$$

for any $N > N_{\epsilon,1}$, where $N_{\epsilon,1}$ is some positive integer depending on ϵ (through $\sigma(\epsilon)$) only. This is equivalent to (93) and completes the feasibility proof. □

The following lemma is needed in our proof of Theorem 2.2.4. We postpone its proof until Section 2.8.

Lemma 2.5.2. *For any $\eta \in (0,1)$ and $\rho_\epsilon \in (0,1)$,*

$$\sum_{j=0}^{\infty} \left[\frac{(N+2\lceil N^\eta\rceil + j - 3)e}{2\lceil N^\eta\rceil - 2}\right]^{2\lceil N^\eta\rceil - 2} \rho_\epsilon^j \leq \frac{(eN)^{2\lceil N^\eta\rceil-2}}{1-\rho_\epsilon}[1+o(1)]. \tag{101}$$

Finally, we provide the proof of Theorem 2.2.4.

Proof of Theorem 2.2.4. Fix any $\epsilon > 0$. Let the sequence of systems be under policy $\pi_\epsilon(N)$ and we consider $N > N_{\epsilon,1}$ as specified by Proposition 2.5.1. Let $Q_{i,\epsilon}^N$, $i = 1, ..., 2\lceil N^\eta\rceil$, be the steady-state total number of type i jobs in the Nth system, and

$$Q_\epsilon^N = \sum_{i=1}^{2\lceil N^\eta\rceil} Q_{i,\epsilon}^N. \tag{102}$$

Using the independence among subsystems, we have that

$$\mathrm{P}\left\{Q_\epsilon^N \geq N\right\} \leq \mathrm{P}\left\{Q_{2\lceil N^\eta\rceil,\epsilon}^N \geq 1\right\} + \mathrm{P}\left\{\sum_{i=1}^{2\lceil N^\eta\rceil-1} Q_{i,\epsilon}^N \geq N-1\right\}. \tag{103}$$

Therefore, if we can prove that

$$\mathrm{P}\left\{Q_{2\lceil N^\eta\rceil,\epsilon}^N \geq 1\right\} \leq e^{-NI(\rho_\epsilon)} \tag{104}$$

and

$$\limsup_{N\to\infty} \frac{1}{N}\log \mathrm{P}\left\{\sum_{i=1}^{2\lceil N^\eta\rceil-1} Q_{i,\epsilon}^N \geq N-1\right\} \leq -I(\rho_\epsilon), \tag{105}$$

then it follows from the principle of the largest term (see Lemma 2.2 in [19]) that

$$\limsup_{N\to\infty} \frac{1}{N}\log \mathrm{P}\left\{Q_\epsilon^N \geq N\right\} \leq -I(\rho_\epsilon). \tag{106}$$

This, together with (12), would establish the weak optimality. In what follows, we shall show (104) and (105).

First, we prove that there exists some $N_{\epsilon,2} \in \mathbb{N}$, whose value only depends on ϵ, such that, for any $N > N_{\epsilon,2}$, (104) holds. Replacing i by $2\lceil N^\eta\rceil - 1$ in (88) yields

$$\begin{aligned} f_{2\lceil N^\eta\rceil-1}(N) =& \lceil N^\eta\rceil^{\alpha^{\lceil N^\eta\rceil-1}} N^{-\gamma(\alpha^{\lceil N^\eta\rceil-2}+2\alpha^{\lceil N^\eta\rceil-3}+\cdots+(\lceil N^\eta\rceil-2)\alpha+(\lceil N^\eta\rceil-1))} \\ \geq& N^{\eta\alpha^{\lceil N^\eta\rceil-1}-\gamma(\alpha^{\lceil N^\eta\rceil-2}+2\alpha^{\lceil N^\eta\rceil-3}+\cdots+(\lceil N^\eta\rceil-2)\alpha+(\lceil N^\eta\rceil-1))} \end{aligned} \tag{107}$$

Using $\alpha > 1$, we obtain the following lower bound for the exponent in (107):

$$\begin{aligned}
&\eta\alpha^{\lceil N^\eta\rceil-1} - \gamma(\alpha^{\lceil N^\eta\rceil-2} + 2\alpha^{\lceil N^\eta\rceil-3} + \cdots + (\lceil N^\eta\rceil - 2)\alpha + (\lceil N^\eta\rceil - 1)) \\
&= \alpha^{\lceil N^\eta\rceil}(\eta\alpha^{-1} + \gamma(-\alpha^{-2} - 2\alpha^{-3} - \cdots - (\lceil N^\eta\rceil - 1)\alpha^{-\lceil N^\eta\rceil})) \\
&\geq \alpha^{\lceil N^\eta\rceil}(\eta\alpha^{-1} + \gamma\sum_{j=2}^{\infty}(-j+1)\alpha^{-j}) \\
&= \alpha^{\lceil N^\eta\rceil}\left(\eta\alpha^{-1} - \frac{\gamma}{(\alpha-1)^2}\right). \qquad (108)
\end{aligned}$$

For convenience, we define

$$C := \eta\alpha^{-1} - \frac{\gamma}{(\alpha-1)^2} = \frac{\eta(\alpha-1)^2 - \gamma\alpha}{\alpha(\alpha-1)^2}, \qquad (109)$$

and it follows from (11) that $C > 0$. Then, combining (107) and (108), we have

$$f_{2\lceil N^\eta\rceil-1}(N) \geq N^{C\alpha^{\lceil N^\eta\rceil}} \geq N^{C\alpha^{N^\eta}}. \qquad (110)$$

With $\mathbf{1}\{A\}$ denoting the indicator function of set A, we obtain the traffic intensity of the $\left(2\lceil N^\eta\rceil\right)$-th subsystem (*i.e.*, the single-server queue processing jobs with their size larger than $r_{2\lceil N^\eta\rceil-1}(N)$) as follows:

$$\begin{aligned}
\rho_{2\lceil N^\eta\rceil}(N) &= N\lambda\cdot E\Big[S\cdot\mathbf{1}\{S > r_{2\lceil N^\eta\rceil-1}(N)\}\Big] = N\lambda\cdot E\Big[S\cdot\mathbf{1}\{S > f_{2\lceil N^\eta\rceil-1}(N)\}\Big] \\
&\leq N\lambda\cdot E[S^\alpha]^{1/\alpha}\cdot P\{S > f_{2\lceil N^\eta\rceil-1}(N)\}^{1-1/\alpha} \qquad (111) \\
&\leq N\lambda\cdot E[S^\alpha]^{1/\alpha}\cdot\left(E[S^\alpha]\Big[f_{2\lceil N^\eta\rceil-1}(N)\Big]^{-\alpha}\right)^{1-1/\alpha} \qquad (112) \\
&\leq N\lambda\cdot E[S^\alpha]\cdot\left(N^{C\alpha^{N^\eta}(\alpha-1)}\right)^{-1} \qquad (113) \\
&= o\Big(e^{-NI(\rho_\epsilon)}\Big), \qquad (114)
\end{aligned}$$

where (111) is due to Hölder's inequality, and (112) and (113) follow from (98) and (110) respectively. Therefore, there exists some $N_{\epsilon,2} \in \mathbb{N}$ such that, for any $N > N_{\epsilon,2}$,

$$\rho_{2\lceil N^\eta\rceil}(N) \leq e^{-NI(\rho_\epsilon)}. \qquad (115)$$

When $\rho_{2\lceil N^\eta\rceil}(N) < 1$, $\mathrm{P}\left\{Q^N_{2\lceil N^\eta\rceil,\epsilon} \geq 1\right\} = \rho_{2\lceil N^\eta\rceil}(N)$. So, for any $N > N_{\epsilon,2}$, (104) holds.

Next we prove (105). Consider those $N > N_\epsilon := \max\{N_{\epsilon,1}, N_{\epsilon,2}\}$. We construct an auxiliary sequence as an upper bound for the original sequence of systems in a similar way as we do in the proof of Theorem 2.2.3. Specifically, for any $N > N_\epsilon$,

- for $i = 1, ..., \lceil N^\eta \rceil$, let all type i jobs have size $\xi_i(N) := s_i(N)\rho_\epsilon \cdot (\lambda^N p_i)^{-1}$,
- for $i = \lceil N^\eta \rceil + 1, ..., 2\lceil N^\eta \rceil - 1$, let all type i jobs have size $\xi_i(N) := s_i(N)\rho_\epsilon \cdot (\lambda^N P_{L_i})^{-1}$,

and everything else remains the same as the original sequence.

As a consequence, the $2\lceil N^\eta \rceil - 1$ independent subsystems in each auxiliary system are all M/D/$\cdot$ queues and they have a common traffic intensity ρ_ϵ. Next, we show that

$$\text{for } i = 1, ..., \lceil N^\eta \rceil, \quad \xi_i(N) \geq f_i(N) = i, \tag{116}$$

$$\text{for } i = \lceil N^\eta \rceil + 1, ..., 2\lceil N^\eta \rceil - 2, \quad \xi_i(N) \geq f_i(N), \tag{117}$$

$$\xi_{2\lceil N^\eta \rceil - 1}(N) \geq f_{2\lceil N^\eta \rceil - 1}(N). \tag{118}$$

To prove (116), for $i = 1, ..., \lceil N^\eta \rceil$, we use (19) to calculate the traffic intensity of the ith subsystem in the Nth original system:

$$\rho_i(N) = \frac{\lambda^N p_i i}{s_i(N)} = \rho_\epsilon \cdot \frac{N p_i i \mu(1 - \sigma(\epsilon))}{\lceil N p_i i \mu(1 - \sigma(\epsilon)) \rceil}, \tag{119}$$

and therefore

$$i = s_i(N)\rho_\epsilon \cdot (\lambda^N p_i)^{-1} \cdot \frac{N p_i i \mu(1 - \sigma(\epsilon))}{\lceil N p_i i \mu(1 - \sigma(\epsilon)) \rceil} \leq s_i(N)\rho_\epsilon \cdot (\lambda^N p_i)^{-1}, \tag{120}$$

which gives (116). We then turn to proving (117). For $i = \lceil N^\eta \rceil + 1, ..., 2\lceil N^\eta \rceil - 2$, the traffic intensity of the ith subsystem in the Nth original system satisfies

$$\rho_i(N) \leq \frac{\lambda^N P_{L_i} f_i(N)}{s_i(N)} = \rho_\epsilon \cdot \frac{N P_{L_i} f_i(N) \mu(1 - \sigma(\epsilon))}{\lceil N P_{L_i} f_i(N) \mu(1 - \sigma(\epsilon)) \rceil} \leq \rho_\epsilon, \tag{121}$$

where the first inequality holds because all type i jobs' sizes in the Nth original system are no greater than $f_i(N)$, and the equality in (121) follows from applying (17). Then (117) immediately follows from (121). To establish (118), we bound the traffic intensity of the $\left(2\lceil N^\eta \rceil - 1\right)$-th subsystem in the Nth original system:

$$\rho_{2\lceil N^\eta \rceil - 1}(N) \leq \frac{\lambda^N P_{L_{2\lceil N^\eta \rceil - 1}} f_{2\lceil N^\eta \rceil - 1}(N)}{s_{2\lceil N^\eta \rceil - 1}(N)} \tag{122}$$

$$\leq \frac{\lambda^N P_{L_{2\lceil N^\eta \rceil - 1}} f_{2\lceil N^\eta \rceil - 1}(N)}{\lceil N P_{L_{2\lceil N^\eta \rceil - 1}} f_{2\lceil N^\eta \rceil - 1}(N) \mu(1 - \sigma(\epsilon)) \rceil} \tag{123}$$

$$= \rho_\epsilon \cdot \frac{N P_{L_{2\lceil N^\eta \rceil - 1}} f_{2\lceil N^\eta \rceil - 1}(N) \mu(1 - \sigma(\epsilon))}{\lceil N P_{L_{2\lceil N^\eta \rceil - 1}} f_{2\lceil N^\eta \rceil - 1}(N) \mu(1 - \sigma(\epsilon)) \rceil}, \tag{124}$$

where (122) is due to the fact that all type $2\lceil N^\eta\rceil-1$ jobs' sizes in the Nth original system are no greater than $f_{2\lceil N^\eta\rceil-1}(N)$, and (123) follows from (92). From (122) $\leq$ (124), we obtain that

$$f_{2\lceil N^\eta\rceil-1}(N) \leq s_{2\lceil N^\eta\rceil-1}(N)\rho_\epsilon \cdot \left(\lambda^N P_{L_{2\lceil N^\eta\rceil-1}}\right)^{-1} \cdot \frac{NP_{L_{2\lceil N^\eta\rceil-1}} f_{2\lceil N^\eta\rceil-1}(N)\mu(1-\sigma(\epsilon))}{\lceil NP_{L_{2\lceil N^\eta\rceil-1}} f_{2\lceil N^\eta\rceil-1}(N)\mu(1-\sigma(\epsilon))\rceil}, \tag{125}$$

which implies (118).

By the construction of the auxiliary sequence, (116), (117) and (118), the only difference between a subsystem in the Nth system of the auxiliary sequence and the corresponding subsystem in the Nth system of the original sequence is that each job in the former can have a greater size (but never a smaller size). Therefore,

$$Q^N_{i,\epsilon} \leq_{\text{st}} Q^N_{i,\epsilon,u}, \quad \text{for all } i = 1, ..., 2\lceil N^\eta\rceil - 1, \tag{126}$$

where $Q^N_{i,\epsilon,u}$ denotes the steady-state total number of type i jobs in the Nth system of the auxiliary sequence, and hence

$$\sum_{i=1}^{2\lceil N^\eta\rceil-1} Q^N_{i,\epsilon} \leq_{\text{st}} \sum_{i=1}^{2\lceil N^\eta\rceil-1} Q^N_{i,\epsilon,u}. \tag{127}$$

Define

$$\overline{s_N} := \Big(s_1(N), ..., s_{2\lceil N^\eta\rceil-1}(N)\Big). \tag{128}$$

Then, for any $N > N_\epsilon$, it follows from (127) that

$$\begin{aligned}
&\mathrm{P}\left\{\sum_{i=1}^{2\lceil N^\eta\rceil-1} Q^N_{i,\epsilon} \geq N-1\right\}\\
\leq &\mathrm{P}\left\{\sum_{i=1}^{2\lceil N^\eta\rceil-1} Q^N_{i,\epsilon,u} \geq N-1\right\}\\
= &\sum_{j=0}^{\infty} \mathrm{P}\left\{\sum_{i=1}^{2\lceil N^\eta\rceil-1} Q^N_{i,\epsilon,u} = N-1+j\right\}\\
\leq &\sum_{j=0}^{\infty} \binom{(N-1+j)+(2\lceil N^\eta\rceil-1)-1}{(2\lceil N^\eta\rceil-1)-1} K(\rho_\epsilon)^{2\lceil N^\eta\rceil-1}\rho_\epsilon^j\\
&\times \mathrm{P}\left\{Q^N_{i,\epsilon,u} \geq s_i(N), \text{for all } i = 1, ..., 2\lceil N^\eta\rceil-1\right\}
\end{aligned} \tag{129}$$

$$\leq \sum_{j=0}^{\infty} \left(\frac{(N + 2\lceil N^\eta \rceil + j - 3)e}{2\lceil N^\eta \rceil - 2} \right)^{2\lceil N^\eta \rceil - 2} K(\rho_\epsilon)^{2\lceil N^\eta \rceil - 1} \rho_\epsilon^j \cdot \prod_{i=1}^{2\lceil N^\eta \rceil - 1} \mathrm{P}\left\{ Q_{i,\epsilon,u}^N \geq s_i(N) \right\} \quad (130)$$

$$\leq \frac{(eN)^{2\lceil N^\eta \rceil - 2}}{1 - \rho_\epsilon} K(\rho_\epsilon)^{2\lceil N^\eta \rceil - 1} \cdot \prod_{i=1}^{2\lceil N^\eta \rceil - 1} \mathrm{P}\left\{ Q_{i,\epsilon,u}^N \geq s_i(N) \right\} [1 + o(1)], \quad (131)$$

where (129) follows from (70) and Lemma 2.4.1, (130) is due to (71) and the independence among $Q_{i,\epsilon,u}^N$'s, and (131) holds by Lemma 2.5.2. Now we apply Lemma 2.3.5 to (131) and obtain the following upper bound

$$\mathrm{P}\left\{ \sum_{i=1}^{2\lceil N^\eta \rceil - 1} Q_{i,\epsilon}^N \geq N - 1 \right\}$$

$$\leq \frac{(eN)^{2\lceil N^\eta \rceil - 2}}{1 - \rho_\epsilon} K(\rho_\epsilon)^{2\lceil N^\eta \rceil - 1} \cdot \frac{e^{-(N-1)I(\rho_\epsilon)}}{[1 - e^{-I(\rho_\epsilon)}]^{2\lceil N^\eta \rceil - 1}} [1 + o(1)], \quad (132)$$

which yields (105), upon taking the logarithm, dividing both sides by N, and letting $N \to \infty$. Finally, combining (104) and (105) yields (106), which together with (12) implies (2). This establishes the weak optimality. □

2.6 General job sizes

In this section, we prove Theorem 2.2.6, which is on systems with general job size distributions.

Proof of Theorem 2.2.6. By the continuity of $I(\cdot)$, (20) and the fact that $\mathrm{E}[S_{\delta_0}] \in \Big[\mathrm{E}[S], \mathrm{E}[S] + \delta_0\Big]$, for any $\epsilon > 0$, there exists $\delta_0 = \delta_0(\epsilon) > 0$ such that (21) holds. Consider a sequence of queues indexed by the number of servers, namely δ_0-systems, with the same parameters as the original sequence in the QD regime, except that the job sizes are equal in distribution to S_{δ_0} as defined in (20). Because S_{δ_0} is a random variable whose possible values have the common divisor δ_0 and $\mathrm{E}[S_{\delta_0}^\alpha] \leq \mathrm{E}[(S + \delta_0)^\alpha] < \infty$, then for any $\epsilon > 0$ we can find a policy $\pi_{\epsilon,g}(N) := \pi_{\epsilon/2}^{\delta_0}(N, S_{\delta_0})$ as given by Corollary 2.2.5 such that

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathrm{P}\{Q_{\epsilon/2,\delta_0}^N \geq N\} < -I(\rho_{\delta_0}) + \frac{1}{2}\epsilon. \quad (133)$$

where $Q_{\epsilon/2,\delta_0}$ denotes the steady-state total number of jobs in the Nth δ_0-system under policy $\pi_{\epsilon,g}(N)$.

For any N, consider the original system with job size S and the δ_0-system with job size S_{δ_0} both under policy $\pi_{\epsilon,g}(N)$. The routing of jobs and the server allocation are the same in these two systems, and their only difference is that some jobs in the δ_0-system have a greater size than the corresponding jobs in the original system. Let Q_ϵ^N denote the steady-state total number of jobs in the original N-server system under policy $\pi_{\epsilon,g}(N)$. From the foregoing argument, it follows that $Q_\epsilon^N \leq_{\text{st}} Q_{\epsilon/2,\delta_0}^N$ and therefore

$$\mathrm{P}\{Q_\epsilon^N \geq N\} \leq \mathrm{P}\{Q_{\epsilon/2,\delta_0}^N \geq N\}. \tag{134}$$

This, together with (133), yields

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathrm{P}\{Q_\epsilon^N \geq N\} < -I(\rho_{\delta_0}) + \frac{1}{2}\epsilon. \tag{135}$$

Finally, combining (135) with (21), we obtain (2) and establish the weak optimality. □

2.7 Concluding remarks

The main operational insight offered by our analysis is that in the QD regime proper size-based job separation can mitigate the impact of job size variability on system performance. This is achieved by incorporating in the policy prescription job size variability level, *i.e.*, its moment index. The proposed SITA policy has the following structure: for small jobs (*i.e.*, type 1 to type $\lceil N^\eta \rceil$), very fine size intervals are divided and load is balanced among these subsystems; for large jobs (*i.e.*, type $\lceil N^\eta \rceil + 1$ to type $2\lceil N^\eta \rceil - 1$), the size intervals become wider and the servers are allocated in such a way that if all jobs sent to each pool attained the maximum value of the corresponding interval, then we would have the workload balanced among these subsystems as well; finally, only one server is reserved for huge jobs (*i.e.*, type $2\lceil N^\eta \rceil$), which are rare due to the finite αth moment assumption on the job size distribution. This structural result provides useful guidance for many computer systems, in which SITA or its variants are selected a priori as the task assignment policy, on how the policy parameters should be chosen in order to reduce the occurrence of congestion.

To the best of our knowledge, this chapter establishes the first analytical result on the steady-state performance of many-server queues with general job size distributions under

any scheduling discipline. The key technical challenge that we address in the performance analysis of the proposed SITA policy $\pi_\epsilon(N)$ in an N-server system is the identification of an auxiliary queueing system that meets three criteria: the auxiliary system (i) serves as a performance upper bound with respect to the chosen metric for the original N-server system under $\pi_\epsilon(N)$, (ii) is easier to analyze than the original system, and (iii) performs sufficiently close to the lower bound system (*i.e.*, the infinite-server queue). Specifically, our upper bound system consists of $[m(N) - 1]$ load-balanced M/D/$s_i(N)$ subsystems, where $m(N) = 2\lceil N^\eta \rceil$ for some $\eta \in (0, 1)$, and a single-server queue. In order to analyze $[m(N) - 1]$ M/D/$s_i(N)$ queues in parallel, we have developed several new results on the M/D/c queue (Theorem 2.3.2, Lemma 2.3.4, Lemma 2.3.5, and Lemma 2.4.1) and utilized existing combinatorial results (*i.e.*, (70) and (71)). In addition, this upper bound system performs ϵ-close— in the sense of Definition 2.2.2— to the infinite-server queue due to our careful construction of the policy, such as the parameterization by the moment index of the job size distribution.

Our main result, however, only states that if the proposed SITA policy is adopted, then weak optimality is achieved. It remains unclear whether such job separation is *necessary*. What if jobs are not separated at all, *e.g.*, is FCFS strongly optimal? In the literature there are asymptotic results on the queue-length process in the G/G/N queue, but the steady-state analysis for G/G/N or even M/G/N queues is still an open problem. Also, what if jobs are separated to a lesser extent than we propose? For example, if we assign all small (or even all small and large) jobs to a single pool instead of $\lceil N^\eta \rceil$ (or $2\lceil N^\eta \rceil - 1$) ones, does the weak optimality result still hold? The answers to these questions seem to depend upon the performance of FCFS in the M/G/c queue with some additional assumptions on the job size distribution, because such policies lead to subsystems in which job sizes are bounded or have a finite support.

Finally, we remark that the variability level of the job size distribution can influence the performance of a scheduling policy. While our recommended SITA policies, by explicitly taking into account the job size variability level through the moment index, achieve weak optimality for both light-tailed and heavy-tailed job sizes, the same may not hold true for

other simpler policies such as FCFS. In fact, the disparity in performance between heavy-tailed and light-tailed cases has been observed in single-server queues (see the discussion in Section 1 of [60] for more details).

2.8 Additional proofs

2.8.1 Proof of Lemma 2.4.1

For any $\overline{r}, \overline{s} \in \mathbb{Z}_+^m$, we define

$$J_1(\overline{r},\overline{s}) := \{i \in \{1,...,m\} : r_i < s_i\},$$

$$J_2(\overline{r},\overline{s}) := \{i \in \{1,...,m\} : r_i > s_i\},$$

$$J_3(\overline{r},\overline{s}) := \{i \in \{1,...,m\} : r_i = s_i\},$$

$$d_k(\overline{r},\overline{s}) := \sum_{i \in J_k(\overline{r},\overline{s})} |r_i - s_i|, \quad \text{for } k = 1, 2.$$

For any $\overline{n} \in S(\overline{c}, j)$, we have

$$\begin{aligned}
&\mathrm{P}\{Q_i = n_i, i = 1,...,m\} \\
&= \mathrm{P}\{Q_i = n_i, i \in J_1(\overline{n},\overline{c})\} \times \mathrm{P}\{Q_i = n_i, i \in J_2(\overline{n},\overline{c})\} \times \mathrm{P}\{Q_i = n_i, i \in J_3(\overline{n},\overline{c})\} \qquad (136) \\
&\leq \rho_o^{-d_1(\overline{n},\overline{c})} \cdot \mathrm{P}\{Q_i = c_i, i \in J_1(\overline{n},\overline{c})\} \times K(\rho_o)^{|J_2(\overline{n},\overline{c})|} \cdot \rho_o^{d_2(\overline{n},\overline{c})} \cdot \mathrm{P}\{Q_i \geq c_i, i \in J_2(\overline{n},\overline{c})\} \\
&\quad \times \mathrm{P}\{Q_i = c_i, i \in J_3(\overline{n},\overline{c})\} \qquad (137) \\
&\leq K(\rho_o)^m \cdot \rho_o^j \cdot \mathrm{P}\{Q_i \geq c_i, \text{ for all } i = 1,...,m\},
\end{aligned}$$

where (136) holds by independence of the m queues and (137) follows from Lemma 2.3.4. Note that, in (137), $|J_2(\overline{n},\overline{c})|$ denotes the cardinality of $J_2(\overline{n},\overline{c})$, which is at most m, and $d_2(\overline{n},\overline{c}) - d_1(\overline{n},\overline{c}) = j$, for any $\overline{n} \in S(\overline{c}, j)$.

2.8.2 Proof of Lemma 2.5.2

$$\begin{aligned}
\sum_{j=0}^{\infty} \left[\frac{(N + 2\lceil N^\eta \rceil + j - 3)e}{2\lceil N^\eta \rceil - 2}\right]^{2\lceil N^\eta \rceil - 2} \rho_\epsilon^j &= \sum_{j=0}^{\infty} e^{2\lceil N^\eta \rceil - 2} \left(1 + \frac{N + j - 1}{2\lceil N^\eta \rceil - 2}\right)^{2\lceil N^\eta \rceil - 2} \rho_\epsilon^j \\
&\leq \sum_{j=0}^{\infty} e^{2\lceil N^\eta \rceil - 2} (N + j)^{2\lceil N^\eta \rceil - 2} \rho_\epsilon^j,
\end{aligned}$$

for all $N \in \mathbb{N}$ such that $2\lceil N^\eta \rceil - 2 \geq 1$. We only consider such N's and show that

$$\sum_{j=0}^{\infty} e^{2\lceil N^\eta \rceil - 2}(N+j)^{2\lceil N^\eta \rceil - 2}\rho_\epsilon^j = \frac{(eN)^{2\lceil N^\eta \rceil - 2}}{1-\rho_\epsilon}[1+o(1)]. \tag{138}$$

For convenience, we define the ratio between the left-hand side and the right-hand side of (138):

$$R := \frac{\sum_{j=0}^{\infty} e^{2\lceil N^\eta \rceil - 2}(N+j)^{2\lceil N^\eta \rceil - 2}\rho_\epsilon^j}{\frac{(eN)^{2\lceil N^\eta \rceil - 2}}{1-\rho_\epsilon}} = \sum_{j=0}^{\infty} \left(1+\frac{j}{N}\right)^{2\lceil N^\eta \rceil - 2} \rho_\epsilon^j(1-\rho_\epsilon), \tag{139}$$

which can be bounded from below and from above as follows

$$1 = \sum_{j=0}^{\infty} \rho_\epsilon^j(1-\rho_\epsilon) \leq R \leq \sum_{j=0}^{\infty} \left(e^{\frac{j}{N}}\right)^{2\lceil N^\eta \rceil - 2} \rho_\epsilon^j(1-\rho_\epsilon) = \sum_{j=0}^{\infty} \left(e^{\frac{2\lceil N^\eta \rceil - 2}{N}}\rho_\epsilon\right)^j (1-\rho_\epsilon). \tag{140}$$

We take N large enough such that $e^{\frac{2\lceil N^\eta \rceil - 2}{N}}\rho_\epsilon < 1$ and then have

$$1 \leq R \leq \frac{1-\rho_\epsilon}{1-e^{\frac{2\lceil N^\eta \rceil - 2}{N}}\rho_\epsilon}.$$

Letting $N \to \infty$ yields the desired limit.

CHAPTER III

REFINING SQUARE-ROOT STAFFING

This chapter is mainly devoted to refining square-root staffing for call centers with impatient customers, modeled by the Erlang A queueing model. We shall also discuss our findings in applying the refined square-root staffing approach to a joint capacity-inventory optimization problem for large-scale manufacturing systems. In what follows, we first focus on the context of call centers and then discuss the capacity-inventory joint optimization problem in Section 3.7.

3.1 Introduction

A key challenge in managing call centers is to balance the trade-off between operational costs and quality-of-service offered to customers. In particular, staffing costs constitute a significant portion of a call center's overall expenditure, which makes it essential to develop adequate models of call center operations that relate operational performance to staffing levels; see [7, 20, 21] for background.

Due to recent theoretical studies, backed up by assessments of empirical data, it is by now widely accepted that the phenomenon of impatient customers (the fact that waiting customers may abandon the system before receiving service) is one of the driving factors for call center performance (see [21] for a thorough discussion). Therefore, multi-server queues with customer abandonments, which explicitly model this phenomenon, have received considerable attention in the literature (see [21, 40], and the references therein). Among different queueing models for call centers with impatient customers, the simplest, yet widely used one is the $M/M/s+M$ model, also referred to as the Erlang A model. Despite the Markovian assumption, this model is considered worthy of being used in practice (see [11, 39]), and its performance analysis has been an important subject of study (see for example [21, 59]).

There is by now a vast literature on the asymptotic analysis of call center models, which

has proven to provide useful managerial insights. In these asymptotic studies, the limiting behavior of a sequence of queues is studied and used to approximate the characteristics of a member of the sequence, i.e., the performance of a finite-sized queueing system. Depending on how this sequence is parameterized, its limiting behavior is different, giving rise to different approximations (see [7, 40]). More specifically, queues with abandonments have been analyzed through fluid approximations (see for example [32, 57, 58, 63]) and diffusion approximations (see, e.g., [15, 38]).

One of the most popular approximations arises in the Quality-and-Efficiency-Driven (QED) regime, in which the number of servers s and the offered workload R are related according to a square-root principle, namely $s = R + \beta\sqrt{R}$, for a constant β. The QED limiting regime for multi-server queues without abandonments were brought to the center of attention by the work of Halfin and Whitt [24]. Garnett et al. [21] study the steady-state performance approximation (as well as a process-level approximation) for the Erlang A model in the QED regime, and Zeltyn and Mandelbaum [62] extend the asymptotic steady-state performance analysis to the $M/M/s+G$ model in the QED regime (as well as in other regimes).

Based on the QED diffusion approximations developed by Halfin and Whitt [24], Borst et al. [7] provide a rigorous justification, in an asymptotic framework, of applying the *square-root staffing* principle to two classes of problems: constraint satisfaction and cost minimization. Here and throughout this chapter, by square-root staffing, we refer to the procedure of calculating the optimal staffing level based on the relevant QED diffusion approximations; this is sometimes simply called QED staffing (see [40]). Borst et al. [7] observe that square-root staffing is accurate over a wide range of system parameters for the Erlang C (or $M/M/s$) model without abandonments. Mandelbaum and Zeltyn [40] apply the results in [62] to the constraint satisfaction problem for the $M/M/s+G$ model, and find that square-root staffing is not as robust as in models without abandonments. In particular, for the Erlang A model, they observe from numerical experiments that square-root staffing is far from optimal for satisfying loose constraints on the tail of the waiting time distribution, and recommend staffing based on a novel limiting approximation for this

particular type of constraint satisfaction problem.

Therefore, for queueing models with abandonments, it is of great interest to understand why the inaccuracy of square-root staffing arises and to develop performance approximations and staffing rules that are accurate in all circumstances. One approach towards accomplishing this goal, which will be taken in this study, is to explicitly characterize, and subsequently correct, the errors of conventional QED diffusion approximation and square-root staffing. Correcting the error of the diffusion approximation, thus obtaining what is known as corrected diffusion approximation, has previously been studied by [5, 47] in the random walk or $GI/G/1$ queue setting and by [29, 30] for the Erlang B (i.e., $M/M/s/s$ loss) and Erlang C models. The explicit characterization of the error of a staffing prescription has received less attention. The only study in this regard is the work by Janssen et al. [30], which develops refined square-root staffing rules for the Erlang C model. In this chapter, we extend their approach to the Erlang A model, a much more realistic model for call centers (see [39]).

Specifically, we consider three different constraint satisfaction problems: (i) delay constraint, which requires the long-run fraction of delayed customers (those not served immediately upon arrival) to be smaller than a certain level; (ii) excess delay constraint, which requires the long-run fraction of customers who wait in the queue for more than T time units, for some $T > 0$, to be lower than some specified level; (iii) abandonment constraint, which requires the long-run fraction of abandoning customers to be smaller than a certain level. In each problem, we search for the lowest staffing level such that the constraint is met. Note that the first two types of constraints are important because they correspond to customers' delay experience, and the third type of constraint is crucial to call centers because customer abandonments result in customer dissatisfaction and potential revenue losses.

Our main results are captured in Theorems 3.3.3, 3.4.2, and 3.5.2, which formally establish the staffing refinement as a characterization of the optimality gap of conventional square-root staffing for each of the three problems. Specifically, our first contribution is to show that as the workload R increases, the difference between the true minimal staffing level that adheres to each constraint, say s_{opt}, and the conventional square-root staffing

prescription, which has the form of $s_* = R + \beta_*\sqrt{R}$ for some (possibly negative) β_*, remains bounded. In fact, $s_{\text{opt}} - s_*$ converges to a real number $\beta_\bullet$ as $R \to \infty$, which refines the existing knowledge (see [40]) that this gap is asymptotically negligible compared to $\sqrt{R}$. Our second and main contribution is to obtain the explicit expression of $\beta_\bullet$ in each case, and prove that the gap between the refined square-root staffing level $s_\bullet = s_* + \beta_\bullet$ and s_{opt} decreases at the rate of $R^{-1/2}$. This provides theoretical support for the improved accuracy of the refined staffing rules. Moreover, the refined rules are as easy to implement as the conventional ones, because the only additional procedure for obtaining $s_\bullet$ is to evaluate and add the refinement $\beta_\bullet$, for which we have explicit expressions. Furthermore, we show that, unlike in the Erlang C model, the refinement $\beta_\bullet$ is significant in many cases, due to different system parameters. One such example is the case of a loose excess delay constraint, which, as mentioned above, was also observed in [40]. But we shall identify more such cases, utilizing the explicit expressions of the refinements that we develop. Our findings suggest that in the presence of customer abandonments, more care needs to be taken in applying many-server asymptotic results to small or moderate size systems, because there are more parameter settings under which it can incur large approximation errors than in the model without abandonments, at least under the Markovian assumption. This also makes the refined staffing rules particularly relevant. We shall demonstrate by numerical experiments that, for most cases of practical interest, refined square-root staffing is very accurate.

Our study provides an analytical assessment of the accuracy of some asymptotic performance approximations and staffing prescriptions. This is a topic on which little research has been done. One related study in this regard is the work by Bassamboo and Randhawa [4], which investigates the accuracy of fluid approximations and the fluid-based capacity prescriptions for cost minimization in the $M/M/s + G$ model. Also, our approach and results potentially can be extended to cost minimization problems with linear cost structures. But if the delay or staffing costs are nonlinear, our approach may not be applicable because conventional square-root staffing may not be asymptotically optimal in that case (see [35]).

The remainder of this chapter is organized as follows. Section 3.2 provides a detailed model description and a technical overview of our refined staffing approach, as well as a

preview of our findings on the influence of abandonments. In Sections 3.3, 3.4, and 3.5, based on corrected diffusion approximations, we develop the refined square-root staffing rules for three constraint satisfaction problems. Section 3.6 contains some concluding remarks. In Section 3.7, we further discuss our findings in applying the refined square-root staffing approach to solving a capacity-inventory joint optimization problem. Finally, various proofs are presented in Section 3.8.

3.2 Model Formulation

In this section, we provide a detailed description of the Erlang A model and our refined square-root staffing approach.

In the Erlang A model, also referred to as the $M/M/s+M$ queue, customers arrive according to a Poisson process with rate λ and require service times that are independent and exponentially distributed with mean $1/\mu$. There are s homogeneous servers working in parallel, and there is unlimited waiting space. Customers that are waiting in the queue abandon the system after an exponentially distributed time with mean $1/\theta$ for some $\theta > 0$. Without loss of generality, we assume $\mu = 1$. Therefore, throughout the remainder of this chapter, the offered workload $R = \lambda$ and the traffic intensity $\rho = \lambda/s$. Let W denote the steady-state waiting time of a customer before receiving service or abandoning the system. The long-run fraction of customers that are not served immediately upon arrival is then given by $P\{W > 0\}$, namely the delay probability. We let $P\{\text{Ab}\}$ denote the long-run fraction of abandoning customers, which can be deduced from the expected steady-state waiting time via the relation $P\{\text{Ab}\} = \theta \cdot E[W]$ (see [39]).

Because of the Markovian assumption, the queue-length process in the Erlang A model is a birth-death process and the exact expressions for $P\{W > T\}$, with $T \geq 0$, and $P\{\text{Ab}\}$ are known (see, e.g., [62] and the appendix of [39]). However, as pointed out by Garnett et al. [21], these expressions are complicated and do not yield much insight. Therefore, we study how these exact expressions behave in the QED asymptotic regime, where λ grows to infinity, the abandonment rate θ does not change with λ, and $s = \lambda + \beta\sqrt{\lambda}$ for some constant β (independent of λ), and more importantly, investigate the implications for staffing. We

next describe this asymptotic approach in detail.

3.2.1 Refined staffing

The core of staffing problems in call centers is to determine the right trade-off between quality and capacity. Quality is formulated in terms of some targeted service level. Take as an example the delay probability $P\{W > 0\}$. A large delay probability is perceived as negative, and the targeted service level could be to keep the delay probability below some value ϵ. The smaller ϵ, the better the offered service. Once the targeted service level is set, the objective from the call center's perspective is to determine the lowest staffing level s such that the target $P\{W > 0\} \leq \epsilon$ is met. This is what we have referred to as a constraint satisfaction problem.

Let us first consider an extension of the delay probability function, namely $A(s, \lambda, \theta)$, allowing the staffing level s to take on any positive real value. Specifically, we define for all $s > 0, \lambda > 0, \theta > 0$,

$$A(s, \lambda, \theta) := \left[1 + \frac{B(s, \lambda)^{-1} - 1}{(s/\theta)e^{\lambda/\theta}(\lambda/\theta)^{-s/\theta}\gamma(s/\theta, \lambda/\theta)}\right]^{-1}, \tag{141}$$

where

$$\gamma(s, a) := \int_0^a t^{s-1}e^{-t}dt, \tag{142}$$

and

$$B(s, \lambda) := \frac{e^{-\lambda}\lambda^s}{\Gamma(s + 1, \lambda)}. \tag{143}$$

Note that for positive integer-valued s, $B(s, \lambda)$ equals the steady-state blocking probability in the $M/M/s/s$ queue (also called the Erlang B formula). In addition, for positive integer-valued s, equation (141) holds if its left-hand side is replaced by the delay probability $P\{W > 0\}$ in the $M/M/s + M$ system (see equation (A.1) in [39]); that is, for positive integer s, the following equation holds

$$P\{W > 0\} = \left[1 + \frac{B(s, \lambda)^{-1} - 1}{(s/\theta)e^{\lambda/\theta}(\lambda/\theta)^{-s/\theta}\gamma(s/\theta, \lambda/\theta)}\right]^{-1}. \tag{144}$$

Hence, (141) indeed defines an extension of $P\{W > 0\}$; i.e., $A(s, \lambda, \theta) = P\{W > 0\}$ for integer-valued s.

In order to determine the lowest staffing level such that $P\{W > 0\} \leq \epsilon$, an exact approach based on the extension function is to first numerically search for the s_{opt} such that $A(s_{\text{opt}}, \lambda, \theta) = \epsilon$ and then round s_{opt} up to the nearest integer. However, as can be seen from (141), the computational complexity of this numerical procedure grows with the magnitude of λ. Also, this exact approach does not yield any operational insight [1].

An alternative approximate approach to solving the delay probability constraint problem is to invoke the theory of asymptotic dimensioning, introduced in [7] and extended in [40] to models with abandonments. Specifically, the performance measures in the Erlang A model can be approximated by their diffusion limit counterparts, e.g., $A(s, \lambda, \theta)$ can be approximated by $A_*(\beta)$, where $A_*(\cdot)$ is some function parameterized by the abandonment rate θ and $\beta := (s - \lambda)\lambda^{-1/2}$. Note that this approximation $A_*(\beta)$ only depends on β and θ (but no longer on the specific value of s or λ). Hence, the inverse problem can be approximatively solved by searching for the β_* such that $A_*(\beta_*) = \epsilon$, calculating $s_* := \lambda + \beta_*\sqrt{\lambda}$, and then setting the staffing level to $\lceil s_* \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. From a computational standpoint, this approximate approach, often called square-root staffing, is much more scalable than the exact one because the hardest part in this procedure, solving for β_*, does not involve λ or s at all. Also, square-root staffing is known to provide useful insights for practitioners (for example, see the discussion on p. 200 to 201 in [52]).

A third approach, recently proposed and applied to the $M/M/s$ model by Janssen et al. [30], refines the square-root staffing approach. Specifically, their goal is also to approximatively solve for s_{opt} such that $C(s_{\text{opt}}, \lambda) = \epsilon$, where $C(s, \lambda)$ denotes an extension of the delay probability function in the $M/M/s$ queue with arrival rate λ and unit service rate. They first prove that for a positive constant β,

$$C(\lambda + \beta\sqrt{\lambda}, \lambda) = C_*(\beta) + C_\bullet(\beta)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{145}$$

where $C_*(\beta)$ denotes the QED diffusion limit, first obtained by Halfin and Whitt as expression (2.3) in [24], and $C_\bullet(\beta)$ is another explicit function of β. Here and throughout

[1] One may also perform the numerical search by evaluating $A(s, \lambda, \theta)$ only at different positive integer s values, but this approach still suffers from computational inscalability and lack of insight. Based on our numerical experience, this method is the slowest.

this chapter, a function $f(\lambda) = \mathcal{O}(g(\lambda))$ if $\limsup_{\lambda\to\infty} |f(\lambda)/g(\lambda)| < \infty$. Then they suggest approximating s_{opt} by

$$s_\bullet := \lambda + \beta_* \sqrt{\lambda} + \beta_\bullet, \tag{146}$$

where β_* solves $C_*(\beta_*) = \epsilon$ and $\beta_\bullet := -C_\bullet(\beta_*)/C'_*(\beta_*)$. Both our and their studies consider the same type of problems, i.e., minimizing the staffing level subject to a service-level constraint, and therefore we can and shall adopt this third approach.

Specifically, in what follows, we first develop approximation results of the same form as (145), namely corrected diffusion approximations, and then uniquely identify $\beta_\bullet$ for the three different constraint satisfaction problems that we introduced in Section 3.1. The refined staffing rules are of the form of (146), with $\beta_\bullet$ some function of β_*, θ, λ, and the constraint target level ϵ that depends on the staffing problem under consideration.

We shall prove that the refined staffing level in (146) yields

$$s_{\text{opt}} - s_\bullet = \mathcal{O}(\lambda^{-1/2}). \tag{147}$$

We refer to the order term that expresses the difference between the exact optimal staffing level and the approximate staffing level as the *optimality gap*. Hence, the optimality gap of $s_\bullet$ is $\mathcal{O}(\lambda^{-1/2})$, which suggests that the staffing level $s_\bullet$ becomes more accurate as λ increases. Note that $s_\bullet = s_* + \beta_\bullet$. We shall also prove that the optimality gap of the conventional staffing level s_* equals $\mathcal{O}(1)$, which indicates that $s_\bullet$ should be a more accurate prescription than s_*. In addition, because $\beta_\bullet$ in fact describes the optimality gap of s_*, or more precisely $s_{\text{opt}} - s_* = \beta_\bullet + \mathcal{O}(\lambda^{-1/2})$, it allows us to perform an analytical assessment of the accuracy of conventional square-root staffing and its underlying QED approximations, and to make some practical recommendations for call center staffing.

3.2.2 The influence of abandonments

Before presenting our results for each constraint satisfaction problem, we briefly summarize the main differences between our findings and those for the Erlang C model in [30], thus highlighting the influence of abandonments.

In the Erlang C model, because other performance measures have simple relations to the delay probability, such as $P\{W > T\} = P\{W > 0\} \cdot e^{-(s-\lambda)T}$ and $E[W] = P\{W >$

$0\}/(s-\lambda)$, Janssen et al. [30] only study one type of constraint satisfaction problem, the delay constraint, in which they find that the optimality gap of square-root staffing is negligible and only becomes slightly larger than one in a few cases. By contrast, for the Erlang A model, we find that for the delay and excess delay constraint problems, due to the presence of customer abandonments, the refinement $\beta_\bullet$ can be quite significant if ϵ, λ, and/or θ are large. For example, for the delay constraint, if θ is very large, $\beta_\bullet$ can be up to nearly 60 (cf. Table 3). Another intriguing observation is that $\beta_\bullet$ is especially significant if the staffing problem leads to an overloaded system, i.e., $\beta_* < 0$ and hence $s_* < \lambda$. For the abandonment constraint problem (which is not applicable to the Erlang C model), $\beta_\bullet$ shows a clear insensitivity to both θ and λ.

3.3 Delay constraint

The objective of the delay constraint satisfaction problem is to determine the number of agents that are required to ensure that $P\{W > 0\}$ is below a threshold ϵ. According to our scheme for refined staffing described in Section 3.2, we shall first derive a corrected diffusion approximation for the objective function, and then solve the asymptotic inverse problem. Throughout this paper, we let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cumulative distribution function and density function, respectively, and further define, for any $\theta > 0$ and $\beta \in (-\infty, \infty)$,

$$G(\beta) = \frac{\Phi(\beta)}{\phi(\beta)}, \quad H_\theta(\beta) = \frac{\phi(\beta/\sqrt{\theta})}{\Phi(-\beta/\sqrt{\theta})}, \tag{148}$$

$$h_\theta(\beta) = -\frac{1}{6}\sqrt{\theta}\beta^2 H_\theta(\beta)\left(G(\beta)H_\theta(\beta)\theta^{-1/2} - \beta G(\beta)\theta^{-1} + 1 + \beta G(\beta)\right). \tag{149}$$

Theorem 3.3.1 (Refined approximation for delay probability)**.** *For any constant $\beta \in (-\infty, \infty)$,*

$$A(\lambda + \beta\sqrt{\lambda}, \lambda, \theta) = A_*(\beta) + A_\bullet(\beta)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{150}$$

where

$$A_*(\beta) = \left(1 + \sqrt{\theta}G(\beta)H_\theta(\beta)\right)^{-1}, \tag{151}$$

$$A_\bullet(\beta) = A_*(\beta)^2\left(\frac{1}{3}\sqrt{\theta}H_\theta(\beta)A_*(\beta)^{-1} - h_\theta(\beta)\right). \tag{152}$$

We emphasize that the dependence on the abandonment rate θ is suppressed in the notation $A_*(\beta)$ and $A_\bullet(\beta)$. Our proof of Theorem 3.3.1 is based on relation (141). First, with $s := \lambda + \beta\sqrt{\lambda}$, a power series approximation in terms of $s^{-1/2}$ is derived for the denominator of the second term in (141), which involves the incomplete gamma function. Then, we combine this result with an approximation of $B(s,\lambda)^{-1}$ developed in [30] to obtain a series approximation of $A(s,\lambda,\theta)^{-1}$ with respect to $s^{-1/2}$. Finally, we derive the desired power series expansion of $A(s,\lambda,\theta)$ in $\lambda^{-1/2}$ using the square-root relation between λ and s. We include the full proof in Section 3.8.1.

The corrected diffusion approximation for the delay probability is thus given by the two terms on the right-hand side of (150), where we ignore the lower order term. If the second term is also ignored, we retrieve the conventional first-order diffusion approximation $A(s,\lambda,\theta) \approx A_*((s-\lambda)\lambda^{-1/2})$, which was derived in [21].

Despite the complicated expression of the corrected diffusion approximation, its computation is as easy as the conventional approximation, because the additional computation of the higher-order term only involves simple algebraic operations on quantities that are already required for evaluating the first-order diffusion approximation (e.g., $G(\beta)$ and $H_\theta(\beta)$).

The second-order refinement term $A_\bullet(\beta)$ turns out to be always positive, which means that the corrected diffusion approximation always takes a larger value, or is more conservative, than the first-order approximation. We state this result, as well as some other properties of the $A_\bullet(\cdot)$ function, as Proposition 3.3.2 and defer its proof until Section 3.8.1.

Proposition 3.3.2. *$A_\bullet(\beta) > 0$ for any $\theta > 0$ and $\beta \in (-\infty,\infty)$, and*

$$\lim_{\beta\to\infty} A_\bullet(\beta) = \lim_{\beta\to-\infty} A_\bullet(\beta) = 0. \tag{153}$$

We shall next use the corrected diffusion approximation to derive a refined staffing level. The refined staffing results (Theorem 3.3.3, and also Theorems 3.4.2 and 3.5.2 in later sections) all follow from the refined performance approximation results (Theorems 3.3.1, 3.4.1, and 3.5.1, respectively) by means of a Taylor expansion argument.

Theorem 3.3.3 (Refined staffing level for delay constraint)**.** *Let $s_{\rm opt} \in (0,\infty)$ be the solution to $A(s_{\rm opt},\lambda,\theta) = \epsilon$, with $\epsilon \in (0,1)$. Let β_* be the solution to $A_*(\beta_*) = \epsilon$, $s_* =$*

$\lambda + \beta_*\sqrt{\lambda}$, *and* $s_\bullet = s_* + \beta_\bullet$ *with*

$$\beta_\bullet = \frac{\beta_*^2}{6}\left(1 - \frac{\sqrt{\theta}H_\theta(\beta_*)}{3h_\theta(\beta_*)\epsilon}\right) > 0\ . \tag{154}$$

Then

$$s_{\text{opt}} - s_* = \mathcal{O}(1), \tag{155}$$

$$s_{\text{opt}} - s_\bullet = \mathcal{O}(\lambda^{-1/2}). \tag{156}$$

Proof. First, a unique β_* exists because $A_*(\beta)$, for $\beta \in (-\infty, \infty)$, decreases from 1 to 0 (see Theorem 4.1 in [40]).

By Proposition 3.3.2, $A_*(\beta_*) + A_\bullet(\beta_*)\lambda^{-1/2} > \epsilon$ and also $A_*(\beta_u) + A_\bullet(\beta_u)\lambda^{-1/2} < \epsilon$ for a sufficiently large β_u. This, together with the continuity of $A_*(\cdot)$ and $A_\bullet(\cdot)$, then implies that there must exist a β_λ such that

$$A_*(\beta_\lambda) + A_\bullet(\beta_\lambda)\lambda^{-1/2} = \epsilon. \tag{157}$$

Let $g(\lambda) := \beta_\lambda - \beta_*$, and then (157) can be rewritten as

$$A_*(\beta_* + g(\lambda)) + A_\bullet(\beta_* + g(\lambda))\lambda^{-1/2} = \epsilon. \tag{158}$$

It follows from the expressions for $A_*(\cdot)$ and $A_\bullet(\cdot)$, i.e., (151) and (152), that their first and second derivative functions are continuous and thus bounded in a small neighborhood of β_*. Therefore, a first-order Taylor expansion of (158) yields

$$A_*(\beta_*) + \mathcal{O}(g(\lambda)) + A_\bullet(\beta_*)\lambda^{-1/2} + \mathcal{O}(g(\lambda)\lambda^{-1/2}) = \epsilon, \tag{159}$$

Also we can further apply a second-order Taylor expansion to (158) to have

$$A_*(\beta_*) + A_*'(\beta_*)g(\lambda) + \mathcal{O}(g(\lambda)^2) + A_\bullet(\beta_*)\lambda^{-1/2} + \mathcal{O}(g(\lambda)\lambda^{-1/2}) = \epsilon. \tag{160}$$

Because $A_*(\beta_*) = \epsilon$ and $A_\bullet(\beta_*) \neq 0$ due to Proposition 3.3.2, it immediately follows from (159) that

$$g(\lambda) = \mathcal{O}(\lambda^{-1/2}). \tag{161}$$

Using (161) and $A_*(\beta_*) = \epsilon$, we solve (160) and obtain that

$$g(\lambda) = -\frac{A_\bullet(\beta_*)}{A_*'(\beta_*)}\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{162}$$

Therefore, β_λ is well approximated by $\beta_* + \beta_\bullet \lambda^{-1/2}$, up to $\mathcal{O}(\lambda^{-1})$, where

$$\beta_\bullet = -\frac{A_\bullet(\beta_*)}{A'_*(\beta_*)}. \tag{163}$$

Because $A_*(\cdot)$ is monotone decreasing (see Theorem 4.1 in [40]), $A'_*(\beta_*) < 0$. Also, $A_\bullet(\beta_*) > 0$ due to Proposition 3.3.2. Therefore, we have $\beta_\bullet > 0$ by (163). By using (151), (152), and $A_*(\beta_*) = \epsilon$, (163) can be further simplified as (154).

We next turn to proving the optimality gap results in (155) and (156). Let $\beta_{\text{opt}} = (s_{\text{opt}} - \lambda)\lambda^{-1/2}$. The desired result is equivalent to

$$\beta_{\text{opt}} - \beta_* = \mathcal{O}(\lambda^{-1/2}), \tag{164}$$

$$\beta_{\text{opt}} - \left(\beta_* + \beta_\bullet \lambda^{-1/2}\right) = \mathcal{O}(\lambda^{-1}). \tag{165}$$

From Theorem 3.3.1, we have that

$$\epsilon = A(\lambda + \beta_{\text{opt}}\sqrt{\lambda}, \lambda, \theta) = A_*(\beta_{\text{opt}}) + \mathcal{O}(\lambda^{-1/2}). \tag{166}$$

Let $g_*(\lambda) := \beta_{\text{opt}} - \beta_*$. Then applying a first-order Taylor expansion to (166), we obtain that

$$\epsilon = A_*(\beta_*) + \mathcal{O}(g_*(\lambda)) + \mathcal{O}(\lambda^{-1/2}). \tag{167}$$

Since $A_*(\beta_*) = \epsilon$, $g_*(\lambda) = \mathcal{O}(\lambda^{-1/2})$, and thus (164) holds. We next prove (165). First, it follows from the derivation of $\beta_\bullet$ that

$$\beta_\lambda - \left(\beta_* + \beta_\bullet \lambda^{-1/2}\right) = \mathcal{O}(\lambda^{-1}). \tag{168}$$

Therefore, in order to conclude (165), it suffices to prove that

$$\beta_{\text{opt}} - \beta_\lambda = \mathcal{O}(\lambda^{-1}). \tag{169}$$

Let $g_\bullet(\lambda) := \beta_{\text{opt}} - \beta_\lambda$. The rest of the proof is similar as above:

$$\epsilon = A(\lambda + \beta_{\text{opt}}\sqrt{\lambda}, \lambda, \theta) = A_*(\beta_{\text{opt}}) + A_\bullet(\beta_{\text{opt}})\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}) \tag{170}$$

$$= A_*(\beta_\lambda) + \mathcal{O}(g_\bullet(\lambda)) + A_\bullet(\beta_\lambda)\lambda^{-1/2} + \mathcal{O}(g_\bullet(\lambda)\lambda^{-1/2}) + \mathcal{O}(\lambda^{-1}). \tag{171}$$

Since $A_*(\beta_\lambda) + A_\bullet(\beta_\lambda)\lambda^{-1/2} = \epsilon$, we find that $g_\bullet(\lambda) = \mathcal{O}(\lambda^{-1})$, which proves the assertion in (169). □

For the delay constraint satisfaction problem, we recommend the refined staffing level $s_\bullet = s_* + \beta_\bullet$, with $\beta_\bullet$ defined in (154). Note that $\beta_\bullet$ is just an explicit function of β_*, θ, and ϵ. Since the classical staffing scheme already requires solving for β_*, which is the hardest task, adapting the refined scheme using $\beta_\bullet$ requires hardly any additional computation. Therefore, we claim that obtaining $s_\bullet$ is as easy as s_*, while $s_\bullet$ achieves a stronger asymptotic optimality than s_*. One interpretation of Theorem 3.3.3 is that $\beta_\bullet$, as defined by (154), exactly captures the dominating term of the error of s_*, or the $\mathcal{O}(1)$ term in (155). By adding the refinement $\beta_\bullet$, the optimality gap of $s_\bullet$ decreases at the rate of $\lambda^{-1/2}$. Also, the fact that $\beta_\bullet$ is always positive suggests that conventional square-root staffing tends to understaff. We remark that it is proved in [40] that $s_{\text{opt}} - s_* = o(\sqrt{\lambda})$, whereas our refined staffing approach enables us to show that the $o(\sqrt{\lambda})$ gap is actually $\mathcal{O}(1)$.

3.3.1 Numerical experiments

We next discuss the numerical experiments that we conducted to illustrate the analytical results. Here, we mainly focus on identifying the scenarios in which $\beta_\bullet$ is large or conventional square-root staffing is not accurate, investigating the accuracy of the refined staffing level $s_\bullet$, and discussing the implications of these findings to call center staffing. Specifically, we vary the values of λ, ϵ, and θ, corresponding to different call center sizes, targeted service levels, and customer patience levels, respectively. Considering call centers of different sizes and targeted service levels is obviously practically relevant. Also it is important to understand the impact of varying customer patience levels as there is not a range of θ values that is widely agreed upon in the literature or in practice. This is partly because different customer patience levels are observed in different call centers (e.g., see [10, 40]). Another reason is that the estimation of θ is quite nontrivial from a methodological standpoint (for example, see Section 7.3 of [11] for a discussion on different estimation methods).

In what follows, we shall show s_{opt}, β_*, s_*, $\beta_\bullet$, and $s_\bullet$ for each problem that we consider. We also include the experimental results on the performance resulting from different staffing levels.

In our extensive numerical experiments, $|s_{\text{opt}} - s_\bullet|$ is almost always less than 1. As

an indication of the error made by the conventional square-root staffing, $\beta_\bullet$ becomes more significant as the abandonment rate θ increases. Also, with the increase of θ, $\beta_\bullet$ gradually becomes a monotone increasing function of the targeted delay probability.

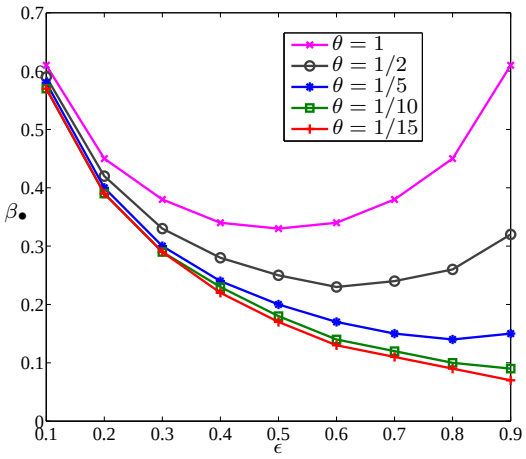

Figure 1: The refinement $\beta_\bullet$ as a function of ϵ, with $\theta \leq 1$.

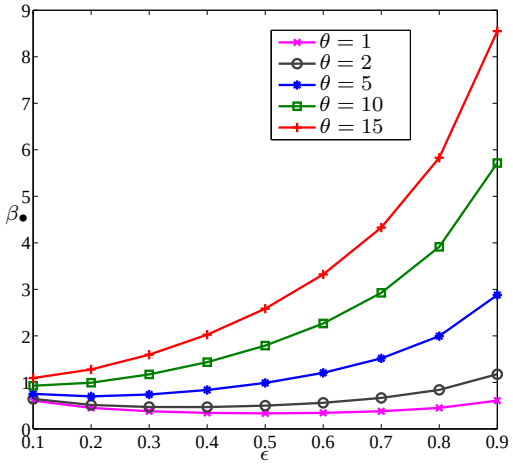

Figure 2: The refinement $\beta_\bullet$ as a function of ϵ, with $\theta \geq 1$.

Figure 1 shows that, when $\theta \leq 1$, $\beta_\bullet$ is always less than 1 and its curve gradually turns to symmetrically bowl-shaped from monotone decreasing in ϵ, as θ increases to 1. In Figure 2, as θ further increases from 1 to 15, $\beta_\bullet$ becomes more significant. In particular, when $\theta \geq 5$, $\beta_\bullet$ is always larger under a looser delay constraint (i.e., a greater ϵ value). For example, as ϵ increases from 0.1 to 0.9, $\beta_\bullet$ increases from about 1 to 6, for $\theta = 10$, and from 1 to nearly

9, for θ=15. Because $\beta_\bullet$ does not depend on λ, such errors are rather severe for a small or moderate size system. For instance, Tables 1 and 2 display the case of $\lambda = 30$, in which the rather large errors are almost completely corrected by $\beta_\bullet$.

Table 1: $P\{W > 0\} = \epsilon, \theta = 10, \lambda = 30$ (high abandonment rate, low call volume)

ϵ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	$A_{\lambda,\theta}(s_*)$	$A_{\lambda,\theta}(s_\bullet)$
0.10	35.64	0.86	34.69	0.93	35.62	0.12	0.10
0.20	32.21	0.22	31.18	0.99	32.18	0.24	0.20
0.30	29.55	-0.30	28.34	1.17	29.51	0.35	0.30
0.40	27.15	-0.79	25.66	1.43	27.10	0.46	0.40
0.50	24.79	-1.29	22.93	1.79	24.72	0.58	0.50
0.60	22.33	-1.83	19.96	2.27	22.23	0.69	0.60
0.70	19.62	-2.46	16.54	2.92	19.46	0.80	0.71
0.80	16.38	-3.25	12.22	3.91	16.13	0.90	0.81
0.90	11.97	-4.43	5.75	5.71	11.46	0.98	0.91

Table 2: $P\{W > 0\} = \epsilon, \theta = 15, \lambda = 30$ (high abandonment rate, low call volume)

ϵ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	$A_{\lambda,\theta}(s_*)$	$A_{\lambda,\theta}(s_\bullet)$
0.10	35.14	0.73	34.02	1.09	35.11	0.13	0.10
0.20	31.51	0.03	30.18	1.28	31.45	0.24	0.20
0.30	28.65	-0.55	26.99	1.59	28.58	0.36	0.30
0.40	26.04	-1.11	23.92	2.02	25.95	0.48	0.40
0.50	23.46	-1.69	20.75	2.58	23.33	0.60	0.50
0.60	20.75	-2.33	17.26	3.32	20.58	0.72	0.61
0.70	17.78	-3.07	13.18	4.33	17.51	0.83	0.71
0.80	14.27	-4.02	7.99	5.83	13.82	0.93	0.81
0.90	9.61	-5.45	0.16	8.55	8.71	1.00	0.92

For large systems, if the customer patience level is low, $\beta_\bullet$ can be quite substantial. For example, Table 3 shows that, when $\theta = 100$, s_* can be off by as many as 20 to 60 servers, while $s_\bullet$ provides an extremely accurate approximation of s_{opt}.

Table 3: $P\{W > 0\} = \epsilon, \theta = 100, \lambda = 3000$ (very high abandonment rate, high call volume)

ϵ	$s_{\rm opt}$	β_*	s_*	$\beta_\bullet$	$s_\bullet$	$A_{\lambda,\theta}(s_*)$	$A_{\lambda,\theta}(s_\bullet)$
0.10	2996.82	-0.12	2993.26	3.51	2996.77	0.11	0.10
0.20	2933.34	-1.32	2927.56	5.70	2933.26	0.21	0.20
0.30	2874.20	-2.45	2865.66	8.42	2874.08	0.31	0.30
0.40	2812.83	-3.63	2800.92	11.75	2812.67	0.42	0.40
0.50	2745.75	-4.94	2729.65	15.87	2745.52	0.52	0.50
0.60	2669.30	-6.43	2647.86	21.11	2668.97	0.63	0.60
0.70	2577.84	-8.23	2549.23	28.12	2577.35	0.73	0.70
0.80	2459.86	-10.58	2420.69	38.37	2459.07	0.83	0.80
0.90	2281.50	-14.18	2223.31	56.72	2280.03	0.92	0.90

In the case of a low customer patience level, or a large θ value, we also compare the square-root staffing prescriptions with those based on the Erlang B model assumption (i.e., $\theta = \infty$). Specifically, we consider two other staffing rules: one based on the exact Erlang B formula $B(s, \lambda)$ (cf. (143)), and the other based on the first-order diffusion approximation for the Erlang B formula, i.e., $B_*(\beta, \lambda) = \phi(\beta)\lambda^{-1/2}/\Phi(\beta)$, where $\beta = (s - \lambda)\lambda^{-1/2}$. Note that the first-order diffusion approximation for the Erlang B formula has the order of $\mathcal{O}(\lambda^{-1/2})$, unlike $\mathcal{O}(1)$ for the Erlang A formula.

As θ becomes very large, compared to the arrival rate λ and the service rate (assumed to be 1), the Erlang A system behaves similarly to the Erlang B model, in which customers have zero patience times, and one expects $A(s, \lambda, \theta) \approx B(s, \lambda) \approx B_*(\beta, \lambda)$. Table 4 shows an example of the comparison, where $s_B := \inf\{s \geq 0 : B(s, \lambda) \leq \epsilon\}$ and $s_B^* := \lambda + \beta_B^*\sqrt{\lambda}$, with $\beta_B^* := \inf\{\beta \in (-\infty, \infty) : B_*(\beta, \lambda) \leq \epsilon\}$. As can be seen in this table, the conventional square-root staffing level s_* is highly biased, while the refined staffing level $s_\bullet$ is extremely accurate for all values of ϵ. It is interesting that both s_B and s_B^* are more accurate than s_*, and the approximative Erlang-B-based staffing level s_B^* turns out to be even slightly better

than the exact one s_B in this case.

Table 4: $P\{W > 0\} = \epsilon, \theta = 100, \lambda = 45$ (very high abandonment rate, low call volume)

ϵ	$s_{\rm opt}$	β_*	s_*	$\beta_\bullet$	$s_\bullet$	s_B	β_B^*	s_B^*	$A_{\lambda,\theta}(s_*)$	$A_{\lambda,\theta}(s_\bullet)$	$A_{\lambda,\theta}(s_B)$	$A_{\lambda,\theta}(s_B^*)$
0.1	47.88	-0.12	44.17	3.51	47.68	45.77	0.21	46.39	0.16	0.10	0.13	0.12
0.2	42.00	-1.32	36.13	5.70	41.83	39.07	-0.77	39.86	0.32	0.20	0.26	0.24
0.3	37.08	-2.45	28.55	8.42	36.97	33.49	-1.59	34.36	0.48	0.30	0.38	0.36
0.4	32.43	-3.63	20.62	11.75	32.37	28.35	-2.35	29.26	0.65	0.40	0.49	0.47
0.5	27.80	-4.94	11.89	15.87	27.76	23.43	-3.08	24.36	0.81	0.50	0.59	0.57
0.6	23.03	-6.43	1.87	21.11	22.98	18.63	-3.79	19.58	0.97	0.60	0.69	0.67
0.7	18.01	-8.23	0.00	28.12	17.91	13.91	-4.49	14.87	1.12	0.70	0.78	0.76
0.8	12.62	-10.58	0.00	38.37	12.42	9.24	-5.19	10.21	1.25	0.80	0.86	0.84
0.9	6.70	-14.18	0.00	56.72	6.59	4.61	-5.88	5.58	1.31	0.90	0.93	0.92

We note that $\beta_\bullet$ tends to be significant when $\beta_* < 0$, as illustrated in Tables 1, 2, 3, and 4. For a number of other cases, especially when $\beta_* > 0$, the refinement $\beta_\bullet$ turns out to be less than one, which provides theoretical support for the adequacy of square-root staffing or QED approximation in those parameter regions. Therefore, we recommend that the refined square-root staffing rule should be adopted for any small to moderate size call center, and for any large size call center with relatively impatient customers, especially if it operates under a moderate or loose delay constraint. In other cases, the conventional staffing rule can be followed without running the risk of substantial inaccuracies.

3.4 Excess delay constraint

We now turn to the constraint satisfaction problem in which the objective function is the steady-state probability that the delay exceeds a certain level T. Specifically, we want to determine the minimum number of agents required to meet the constraint $P\{W > T\} \leq \epsilon$. We start by deriving a corrected diffusion approximation for $A_t(s, \lambda, \theta) := P\{W > t\lambda^{-1/2}\}$.

Theorem 3.4.1 (Refined approximation for excess delay)**.** *For any constant $\beta \in (-\infty, \infty)$,*

$$A_t(\lambda + \beta\sqrt{\lambda}, \lambda, \theta) = A_*(\beta)d_*(\beta, t) + \left[A_*(\beta)d_\bullet(\beta, t) + A_\bullet(\beta)d_*(\beta, t)\right]\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \quad (172)$$

where

$$d_*(\beta, t) = \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\Phi(-\beta\theta^{-1/2})}, \quad (173)$$

$$d_\bullet(\beta,t) = d_*(\beta,t)\left(\frac{1}{6}I_\bullet(\beta,\theta/2,t)\frac{\theta^{5/2}\phi(\beta\theta^{-1/2})}{\Phi(-\sqrt{\theta}t-\beta\theta^{-1/2})} - \frac{1}{6}I_\bullet(\beta,\theta/2,0)\theta^{5/2}H_\theta(\beta) - \theta t\right), \tag{174}$$

$$I_\bullet(a,b,t) = \int_t^\infty \exp\{-ay-by^2\}y^3dy, \quad \forall a>0, b>0, t\geq 0. \tag{175}$$

The main step in the proof of Theorem 3.4.1 is to show that

$$P\{W > t\lambda^{-1/2}|W>0\} = d_*(\beta,t) + d_\bullet(\beta,t)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{176}$$

We prove (176) by deriving and combining corrected approximations for the two integral-form building blocks of the exact expression for $P\{W > t\lambda^{-1/2}|W>0\}$. In particular, we apply the Laplace method to analyze their asymptotic behavior and refine the results presented in Section 10 and Theorem 4.1(g) in [62]. The detailed proof is included in Section 3.8.2.

The right-hand side of (172), excluding the order term, serves as the corrected diffusion approximation for $P\{W > t\lambda^{-1/2}\}$, while the conventional diffusion approximation is given by the first term only, i.e., $P\{W > t\lambda^{-1/2}\} \approx A_*(\beta)d_*(\beta,t)$. Again, the evaluation of the correction term only involves simple algebra on known quantities from the computation of the conventional diffusion approximation, and in particular $I_\bullet(a,b,t)$ can be calculated fast using (250), where it is expressed explicitly in terms of the standard normal distribution function.

Now we first consider the constraint of the form $P\{W > t\lambda^{-1/2}\} \leq \epsilon$. Because the (corrected) diffusion approximations for $P\{W > t\lambda^{-1/2}\}$ in (172) and $P\{W>0\}$ in (150) have exactly the same order in each corresponding term, the staffing procedure in Section 3.3 and, in particular, the expression (163) can be directly applied here with proper substitutions, leading to the following result:

Theorem 3.4.2 (Refined staffing level for excess delay constraint)**.** *Let $s_{\rm opt} \in (0,\infty)$ be the solution to $A_t(s_{\rm opt},\lambda,\theta) = \epsilon$, for some $t>0$* and *$\epsilon \in (0,1)$. Let β_* be the solution to $A_*(\beta_*)d_*(\beta_*,t) = \epsilon$, $s_* = \lambda + \beta_*\sqrt{\lambda}$, and $s_\bullet = s_* + \beta_\bullet$ with*

$$\beta_\bullet = -\frac{A_*(\beta_*)d_\bullet(\beta_*,t) + A_\bullet(\beta_*)d_*(\beta_*,t)}{A'_*(\beta_*)d_*(\beta_*,t) + A_*(\beta_*)d'_*(\beta_*,t)}, \tag{177}$$

where $d'_*(\cdot,\cdot)$ denotes the derivative of $d_*(\cdot,\cdot)$ with respect to the first argument. Then,

$$s_{\text{opt}} - s_* = \mathcal{O}(1), \tag{178}$$

$$s_{\text{opt}} - s_\bullet = \mathcal{O}(\lambda^{-1/2}). \tag{179}$$

We note that because the second-order term of the corrected diffusion approximation in this case $A_*(\beta)d_\bullet(\beta,t) + A_\bullet(\beta)d_*(\beta,t)$ is not always positive, the staffing refinement $\beta_\bullet$ may be negative or zero as well (see Table 6 for an example). If $\beta_\bullet = 0$, the optimality gap is $\mathcal{O}(\lambda^{-1/2})$ for both s_* and $s_\bullet$, which are equal. The proof of Theorem 3.4.2 can be found in Section 3.8.2.

For staffing in practice, when the constraint has the form $P\{W > T\} \leq \epsilon$, for a fixed T, we let $t = T\sqrt{\lambda}$. Then the constraint to satisfy becomes $P\{W > t\lambda^{-1/2}\} \leq \epsilon$, and the above staffing rule applies. In this case, $\beta_\bullet$ depends on θ, ϵ, λ, and T (through β_* and t).

3.4.1 Numerical experiments

In this subsection, we investigate numerically the gain of refined staffing. We also compare square-root staffing, both conventional and refined, with ED+QED staffing, which is a staffing principle developed for satisfying the excess delay constraint in [40]. Specifically, for the constraint $P\{W > T\} \leq \epsilon$, Theorem 4.4 in [40] prescribes the staffing level

$$s_{\text{EQ}} = e^{-\theta T}\lambda + \delta^*\sqrt{\lambda}, \tag{180}$$

where

$$\delta^* = \Phi^{-1}(1 - \epsilon \cdot e^{\theta T})\sqrt{\theta e^{-\theta T}}. \tag{181}$$

Note that, if $\epsilon \geq e^{-\theta T}$, $s_{\text{opt}} = 0$, because with zero server a customer's waiting time is just his patience time and thus setting $s = 0$ yields $P\{W > T\} = e^{-\theta T} \leq \epsilon$. We do not consider such cases.

First, we focus on the constraints with small T values, which describes some of the key performance measures for call centers. For example, extremely small T and ϵ values may correspond to emergency call centers, such as 911 in the U.S., and $P\{W >$20 seconds$\} \leq \epsilon$, for some ϵ at the order of 10%, is the rule of thumb for many other types of call centers.

Table 5: $P\{W > 0.05\} = \epsilon, \theta = 0.5, \lambda = 30, \epsilon = 0.001$ to 0.01 (low abandonment rate, low call volume, tight constraints)

ϵ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	s_{EQ}	$P\{W > 0.05\}\|_{s=s_*}$	$P\{W > 0.05\}\|_{s=s_\bullet}$	$P\{W > 0.05\}\|_{s=s_{\text{EQ}}}$
0.001	47.00	2.85	45.58	1.50	47.09	41.05	0.0021	0.0010	0.0177
0.002	45.69	2.64	44.44	1.32	45.76	40.24	0.0037	0.0019	0.0247
0.003	44.89	2.51	43.75	1.21	44.95	39.74	0.0052	0.0029	0.0302
0.004	44.31	2.42	43.23	1.13	44.37	39.37	0.0067	0.0039	0.0349
0.005	43.85	2.34	42.83	1.08	43.90	39.08	0.0081	0.0049	0.0391
0.006	43.46	2.28	42.49	1.03	43.52	38.83	0.0094	0.0059	0.0429
0.007	43.13	2.23	42.19	0.99	43.18	38.62	0.0107	0.0068	0.0464
0.008	42.84	2.18	41.94	0.96	42.89	38.44	0.0120	0.0078	0.0497
0.009	42.59	2.14	41.71	0.93	42.63	38.27	0.0133	0.0088	0.0529
0.010	42.35	2.10	41.50	0.90	42.40	38.12	0.0146	0.0098	0.0558

Table 6: $P\{W > 0.05\} = \epsilon, \theta = 0.5, \lambda = 30, \epsilon = 0.1$ to 0.9 (low abandonment rate, low call volume, moderate to loose constraints)

ϵ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	s_{EQ}	$P\{W > 0.05\}\|_{s=s_*}$	$P\{W > 0.05\}\|_{s=s_\bullet}$	$P\{W > 0.05\}\|_{s=s_{\text{EQ}}}$
0.1	36.43	1.11	36.08	0.37	36.45	34.11	0.1119	0.0995	0.2006
0.2	34.12	0.71	33.90	0.23	34.13	32.41	0.2123	0.1994	0.3084
0.3	32.53	0.43	32.37	0.16	32.54	31.18	0.3109	0.2995	0.4030
0.4	31.22	0.20	31.11	0.12	31.22	30.13	0.4091	0.3995	0.4919
0.5	30.04	-0.01	29.95	0.09	30.04	29.14	0.5071	0.4997	0.5782
0.6	28.89	-0.21	28.83	0.06	28.89	28.14	0.6051	0.5998	0.6632
0.7	27.69	-0.43	27.65	0.04	27.69	27.06	0.7029	0.7000	0.7481
0.8	26.30	-0.67	26.30	-0.00	26.30	25.75	0.7999	0.8001	0.8331
0.9	24.34	-1.01	24.49	-0.14	24.35	23.81	0.8942	0.8994	0.9178

Table 7: $P\{W > 0.05\} = \epsilon, \theta = 4, \lambda = 30$ (high abandonment rate, low call volume, tight constraints)

ϵ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	s_{EQ}	$P\{W > 0.05\}\|_{s=s_*}$	$P\{W > 0.05\}\|_{s=s_\bullet}$	$P\{W > 0.05\}\|_{s=s_{\text{EQ}}}$
0.001	45.79	2.68	44.68	1.23	45.90	54.60	0.0017	0.0009	0.0000
0.002	44.36	2.45	43.43	1.03	44.46	52.46	0.0031	0.0019	0.0000
0.003	43.48	2.31	42.65	0.92	43.57	51.14	0.0043	0.0029	0.0001
0.004	42.83	2.21	42.08	0.84	42.92	50.17	0.0055	0.0039	0.0001
0.005	42.31	2.12	41.62	0.78	42.40	49.40	0.0067	0.0048	0.0001
0.006	41.88	2.05	41.23	0.73	41.96	48.75	0.0078	0.0058	0.0002
0.007	41.51	1.99	40.90	0.68	41.58	48.20	0.0089	0.0068	0.0003
0.008	41.18	1.94	40.60	0.65	41.25	47.71	0.0100	0.0078	0.0004
0.009	40.88	1.89	40.34	0.62	40.95	47.27	0.0111	0.0087	0.0005
0.010	40.61	1.84	40.09	0.59	40.68	46.87	0.0122	0.0097	0.0006

In this case, if the abandonment rate is low, the conventional square-root staffing is extremely accurate, regardless of the system size or the targeted service level. Tables 5 and 6 illustrate the cases for small λ values; similar findings hold for other λ and ϵ values. ED+QED staffing tends to prescribe staffing levels that are too low, especially under tight constraints, as shown in Table 5. This parameter region is of particular interest to the staffing of emergency call centers, having relatively patient customers and tight delay constraints.

Table 8: $P\{W > 0.05\} = \epsilon, \theta = 4, \lambda = 1000$ (high abandonment rate, low call volume, moderate to loose constraints)

ϵ	$s_{\rm opt}$	β_*	s_*	$\beta_\bullet$	$s_\bullet$	$s_{\rm EQ}$	$P\{W > 0.05\}\|_{s=s_*}$	$P\{W > 0.05\}\|_{s=s_\bullet}$	$P\{W > 0.05\}\|_{s=s_{\rm EQ}}$
0.05	909.68	-3.05	903.68	6.53	910.22	907.20	0.061	0.049	0.054
0.10	887.41	-3.77	880.91	7.09	887.99	885.36	0.120	0.098	0.106
0.15	872.19	-4.25	865.47	7.34	872.81	870.42	0.176	0.148	0.157
0.20	859.96	-4.64	853.18	7.44	860.62	858.37	0.231	0.197	0.207
0.25	849.33	-4.98	842.63	7.41	850.04	847.86	0.284	0.246	0.257
0.30	839.65	-5.28	833.15	7.28	840.43	838.27	0.335	0.296	0.307
0.35	830.53	-5.55	824.36	7.04	831.40	829.19	0.385	0.345	0.358
0.40	821.70	-5.82	816.01	6.68	822.69	820.37	0.432	0.394	0.408
0.45	812.93	-6.07	807.94	6.17	814.11	811.59	0.478	0.443	0.458
0.50	804.03	-6.32	800.00	5.48	805.48	802.64	0.522	0.492	0.508

If the abandonment rate is high, the conventional square-root staffing is still very accurate for small systems (or small λ's), while ED+QED staffing tends to overstaff, especially under tight constraints (see Table 7). For large λ's, when the constraint can be satisfied with the system being overloaded, $\beta_\bullet$ becomes substantial and $s_{\rm EQ}$ also becomes more accurate than s_*. Table 8 shows such an example.

Next, we consider the constraints with moderate or large T values. As illustrated in [40], s_* is accurate when the load is small, but not so when the load is moderate or large. In the latter case, the refinement significantly improves the accuracy. Table 9 displays the same example as considered in Section 5.3 of the online appendix of [40]. For $P\{W > \frac{1}{3}\}$, $\theta = 0.5$, and $\lambda = 1000$, s_* always underestimates $s_{\rm opt}$ by nearly 10 servers, while the difference between $s_{\rm opt}$ and $s_\bullet$ is less than 1. Note that in this case $s_{\rm EQ}$ is slightly more accurate than $s_\bullet$, which suggests that the ED+QED regime is better modeled as an overloaded regime.

Table 9: $P\{W > \frac{1}{3}\} = \epsilon, \theta = 0.5, \lambda = 1000$ (low abandonment rate, high call volume, moderate to loose constraints)

ϵ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	s_{EQ}	$P\{W > \frac{1}{3}\}\|_{s=s_*}$	$P\{W > \frac{1}{3}\}\|_{s=s_\bullet}$	$P\{W > \frac{1}{3}\}\|_{s=s_{\text{EQ}}}$
0.05	879.00	-4.11	870.11	9.41	879.52	878.63	0.108	0.048	0.052
0.10	871.13	-4.36	861.99	9.68	871.67	870.85	0.194	0.096	0.102
0.15	865.77	-4.54	856.51	9.82	866.32	865.53	0.268	0.144	0.153
0.20	861.47	-4.68	852.15	9.88	862.04	861.26	0.334	0.193	0.203
0.25	857.74	-4.79	848.42	9.90	858.32	857.55	0.394	0.242	0.253
0.30	854.34	-4.90	845.06	9.89	854.95	854.16	0.449	0.291	0.303
0.35	851.15	-5.00	841.95	9.83	851.78	850.98	0.500	0.340	0.353
0.40	848.07	-5.09	839.00	9.73	848.73	847.90	0.546	0.389	0.403
0.45	845.02	-5.18	836.14	9.59	845.73	844.85	0.589	0.438	0.453
0.50	841.94	-5.27	833.33	9.38	842.72	841.76	0.628	0.487	0.503

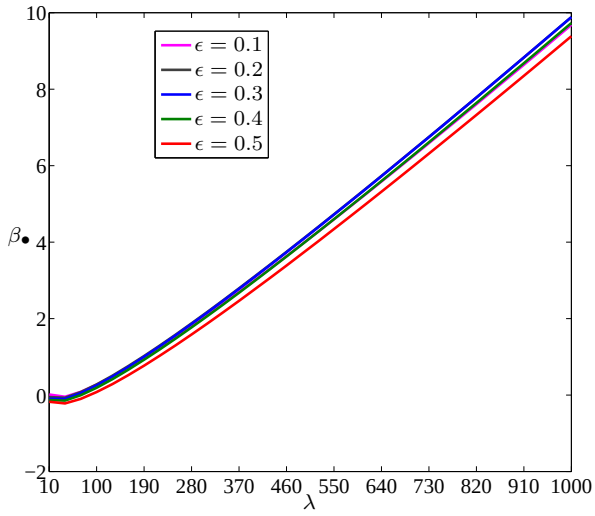

Figure 3: The refinement $\beta_\bullet$ as a function of λ, for $P\{W > \frac{1}{3}\} = \epsilon$ with $\theta = 0.5$. The five lines corresponding to different ϵ values are either indistinguishable or very close.

The fact that s_*, as an asymptotic approximation, is less accurate for larger λ values might seem counterintuitive, but it can be easily explained with the aid of the explicit $\beta_\bullet$ expression. Again, we consider the above example, i.e., $T = \frac{1}{3}$ and $\theta = 0.5$. In Figure 3, with ϵ fixed at different values, we plot the $\beta_\bullet$, as a function of λ, calculated by (177). The plot clearly shows the growth of $\beta_\bullet$ with λ. It is interesting to note that the increase is approximately linear and that the five lines corresponding to different ϵ values do not differ much. The explanation of the seeming discrepancy between this increase of $\beta_\bullet$ and Theorem

3.4.2 is as follows. The asymptotic optimality results for s_* and $s_\bullet$ hold for the constraint of the form $P\{W > t\lambda^{-1/2}\} \leq \epsilon$, where t does not scale with λ. In this example, as we vary λ while fixing $t\lambda^{-1/2}$ at $\frac{1}{3}$, the value of t changes as well, and therefore Figure 3 should not be considered conflicting with the asymptotic optimality results stated in Theorem 3.4.2.

In summary, for the excess delay constraint satisfaction problem, we recommend that refined staffing should always be adopted. Also, the experimental results show that the accuracy improvement due to the refinement is especially significant if $\beta_* < 0$; this is the same as in Section 3.3.

3.5 Abandonment constraint

In this section, we develop the refined staffing rule for satisfying the constraint on the steady-state abandonment probability. Again, we start with a refined diffusion approximation for $b(s, \lambda, \theta) := P\{\text{Ab}\}$.

Theorem 3.5.1 (Refined approximation for abandonment probability)**.** *For any constant* $\beta \in (-\infty, \infty)$,

$$b(\lambda + \beta\sqrt{\lambda}, \lambda, \theta) = b_*(\beta)\lambda^{-1/2} + b_\bullet(\beta)\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}), \tag{182}$$

where

$$b_*(\beta) = (\sqrt{\theta}H_\theta(\beta) - \beta)A_*(\beta), \qquad b_\bullet(\beta) = u_\theta(\beta)b_*(\beta), \tag{183}$$

$$u_\theta(\beta) = -h_\theta(\beta)A_*(\beta) - \frac{1}{6}\beta^2 H_\theta(\beta)\theta^{-1/2} + \frac{1}{6}\beta H_\theta(\beta)\sqrt{\theta}\Big(\sqrt{\theta}H_\theta(\beta) - \beta\Big)^{-1}. \tag{184}$$

We prove Theorem 3.5.1 by first deriving a power series approximation of $P\{\text{Ab}|W > 0\}$ in terms of $s^{-1/2}$, then combining this with the refined approximation of $P\{W > 0\}$ to get the series expansion of $P\{\text{Ab}\}$ in terms of $s^{-1/2}$, and finally obtaining (182) by exploiting the square-root relation between λ and s. The full proof can be found in Section 3.8.3.

To consider the abandonment constraint problem, we start by deriving the refined staffing rule and proving its stronger asymptotic optimality for the constraint of the form $P\{\text{Ab}\} \leq \epsilon\lambda^{-1/2}$. Then we discuss how to apply the refined staffing rule to solving the

abandonment constraint problem of the unscaled form $P\{\text{Ab}\} \leq \epsilon$, and present our numerical results. This follows the same procedure as the conventional square-root staffing in [40] (see their expression (19) and Remark 4.3).

Theorem 3.5.2 (Refined staffing level for abandonment constraint)**.** *Let* $s_{\text{opt}} \in (0, \infty)$ *be the solution to* $b(s_{\text{opt}}, \lambda, \theta) = \epsilon\lambda^{-1/2}$, with $\epsilon\lambda^{-1/2} \in (0, 1)$. *Let* β_* *be the solution to* $b_*(\beta_*)\lambda^{-1/2} = \epsilon\lambda^{-1/2}$ *or* $b_*(\beta) = \epsilon$, $s_* = \lambda + \beta_*\sqrt{\lambda}$, *and* $s_\bullet = s_* + \beta_\bullet$ *with*

$$\beta_\bullet = -\frac{b_\bullet(\beta_*)}{b'_*(\beta_*)}. \tag{185}$$

Then,

$$s_{\text{opt}} - s_* = \mathcal{O}(1), \tag{186}$$

$$s_{\text{opt}} - s_\bullet = \mathcal{O}(\lambda^{-1/2}). \tag{187}$$

The proof of Theorem 3.5.2 is very similar to that of Theorem 3.3.3 and thus is omitted. Furthermore, simple calculations show that

$$b_\bullet(\beta_*) = u_\theta(\beta_*)\epsilon \tag{188}$$

and

$$b'_*(\beta_*) = \left(6A_*(\beta_*)h_\theta(\beta_*)\beta_*^{-2} - \beta_*\theta^{-1}\right)\epsilon + \left(H_\theta(\beta_*)^2 - \beta_*^2\theta^{-1} - 1\right)A_*(\beta_*). \tag{189}$$

Therefore, one may use (188) and (189) to evaluate (185). Also, similar to the excess delay constraint problem, $\beta_\bullet$ can be negative or zero in this case (see Tables 14 and 15).

The abandonment constraint in call center practice has the form $P\{\text{Ab}\} \leq \epsilon$. In this case we should first solve for β_* such that $b_*(\beta_*) = \epsilon\sqrt{\lambda}$ (i.e., the conventional square-root staffing procedure as suggested in Remark 4.3 of [40]), and then calculate the refinement (185) with the solution β_*. Although the asymptotic optimality results only hold for the problem stated in Theorem 3.5.2, our numerical experiments will show that the staffing refinement captures the error of the conventional square-root staffing prescription for satisfying $P\{\text{Ab}\} \leq \epsilon$ as well.

3.5.1 Numerical experiments

In this subsection, we present some numerical results for satisfying the constraint $P\{\text{Ab}\} \leq \epsilon$. We follow the translation procedure described above, i.e., solving $b_*(\beta_*) = \epsilon\sqrt{\lambda}$ and then calculating the refinement (185) with the solution β_*. When the abandonment probability constraint becomes very tight (ϵ= 0.1% or even smaller), $\beta_\bullet$ becomes non-negligible and its magnitude is not sensitive to the abandonment rate or the offered load. For example, Tables 10 and 11 show that, for $\epsilon = 10^{-5}$, s_* is always off by a couple of servers, for a wide range of θ and λ values.

Table 10: $P\{\text{Ab}\} = \epsilon$, with $\epsilon = 10^{-5}$ and $\theta = 1$ (low abandonment rate, tight constraint)

λ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	$P\{\text{Ab}\}\|_{s=s_*}$	$P\{\text{Ab}\}\|_{s=s_\bullet}$
10	23.70	3.65	21.55	2.37	23.92	7.35×10^{-5}	8.06×10^{-6}
20	38.06	3.57	35.95	2.27	38.22	4.39×10^{-5}	8.91×10^{-6}
50	76.44	3.45	74.41	2.13	76.54	2.60×10^{-5}	9.53×10^{-6}
100	135.59	3.36	133.63	2.03	135.66	1.95×10^{-5}	9.76×10^{-6}
200	248.16	3.27	246.28	1.93	248.20	1.59×10^{-5}	9.89×10^{-6}
500	572.18	3.15	570.41	1.80	572.21	1.32×10^{-5}	9.96×10^{-6}
1000	1098.23	3.05	1096.55	1.70	1098.25	1.20×10^{-5}	9.98×10^{-6}

Table 11: $P\{\text{Ab}\} = \epsilon$, with $\epsilon = 10^{-5}$ and $\theta = 50$ (high abandonment rate, tight constraint)

λ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	$P\{\text{Ab}\}\|_{s=s_*}$	$P\{\text{Ab}\}\|_{s=s_\bullet}$
10	25.86	4.19	23.24	3.08	26.33	1.14×10^{-4}	6.27×10^{-6}
20	40.99	4.11	38.37	2.97	41.34	6.44×10^{-5}	7.70×10^{-6}
50	80.87	4.00	78.27	2.83	81.10	3.50×10^{-5}	8.93×10^{-6}
100	141.69	3.91	139.14	2.71	141.85	2.46×10^{-5}	9.44×10^{-6}
200	256.62	3.83	254.13	2.60	256.73	1.88×10^{-5}	9.72×10^{-6}
500	585.36	3.71	582.97	2.45	585.42	1.47×10^{-5}	9.89×10^{-6}
1000	1116.76	3.62	1114.46	2.34	1116.81	1.30×10^{-5}	9.95×10^{-6}

For loose or moderate constraints, $|\beta_\bullet|$ is less than 1 in most cases. Tables 12 and 13 display examples for moderate constraints, and Tables 14 and 15 illustrate the loose constraint case. We observe that, unlike in the other two types of problems, the abandonment rate θ does not affect the magnitude of $\beta_\bullet$ much and conventional square-root staffing does not become inaccurate, and in fact is still quite accurate, when the constraint leads to an overloaded system (see Tables 14 and 15). Again, in all cases, the refined square-root staffing rule yields an accurate approximation of s_{opt}. Therefore, we recommend that, for call centers with a tight abandonment constraint, the refined staffing procedure should be followed, regardless of the customer patience level, and s_* can be used otherwise.

Table 12: $P\{\text{Ab}\} = \epsilon$, with $\epsilon = 10^{-2}$ and $\theta = 1$ (low abandonment rate, moderate constraint)

λ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	$P\{\text{Ab}\}\|_{s=s_*}$	$P\{\text{Ab}\}\|_{s=s_\bullet}$
10	15.06	1.47	14.64	0.47	15.10	1.289×10^{-2}	9.704×10^{-3}
20	26.21	1.31	25.85	0.39	26.24	1.165×10^{-2}	9.864×10^{-3}
50	57.94	1.08	57.67	0.29	57.96	1.073×10^{-2}	9.954×10^{-3}
100	109.23	0.90	109.02	0.22	109.24	1.037×10^{-2}	9.980×10^{-3}
200	210.13	0.71	209.99	0.15	210.14	1.017×10^{-2}	9.992×10^{-3}
500	509.46	0.42	509.39	0.08	509.47	1.005×10^{-2}	9.998×10^{-3}
1000	1005.65	0.18	1005.63	0.03	1005.66	1.001×10^{-2}	9.999×10^{-3}

Table 13: $P\{\text{Ab}\} = \epsilon$, with $\epsilon = 10^{-2}$ and $\theta = 50$ (high abandonment rate, moderate constraint)

λ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	$P\{\text{Ab}\}\|_{s=s_*}$	$P\{\text{Ab}\}\|_{s=s_\bullet}$
10	17.23	2.09	16.62	0.88	17.51	1.421×10^{-2}	8.454×10^{-3}
20	29.24	1.94	28.67	0.78	29.44	1.265×10^{-2}	9.166×10^{-3}
50	62.63	1.71	62.12	0.64	62.76	1.136×10^{-2}	9.676×10^{-3}
100	115.73	1.53	115.28	0.53	115.82	1.078×10^{-2}	9.851×10^{-3}
200	219.12	1.33	218.75	0.43	219.18	1.042×10^{-2}	9.934×10^{-3}
500	523.16	1.02	522.90	0.30	523.20	1.016×10^{-2}	9.978×10^{-3}
1000	1024.30	0.76	1024.12	0.20	1024.32	1.007×10^{-2}	9.991×10^{-3}

Table 14: $P\{\text{Ab}\} = \epsilon$, with $\epsilon = 0.2$ and $\theta = 1$ (low abandonment rate, loose constraint)

λ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	$P\{\text{Ab}\}\|_{s=s_*}$	$P\{\text{Ab}\}\|_{s=s_\bullet}$
10	8.67	-0.40	8.73	-0.04	8.69	1.964×10^{-1}	1.988×10^{-1}
20	16.51	-0.77	16.57	-0.05	16.52	1.977×10^{-1}	1.996×10^{-1}
50	40.23	-1.38	40.27	-0.04	40.23	1.992×10^{-1}	2.000×10^{-1}
100	80.07	-1.99	80.09	-0.02	80.07	1.998×10^{-1}	2.000×10^{-1}
200	160.01	-2.83	160.01	-0.00	160.01	2.000×10^{-1}	2.000×10^{-1}
500	400.00	-4.47	400.00	-0.00	400.00	2.000×10^{-1}	2.000×10^{-1}
1000	800.00	-6.32	800.00	-0.00	800.00	2.000×10^{-1}	2.000×10^{-1}

Table 15: $P\{\text{Ab}\} = \epsilon$, with $\epsilon = 0.2$ and $\theta = 50$ (high abandonment rate, loose constraint)

λ	s_{opt}	β_*	s_*	$\beta_\bullet$	$s_\bullet$	$P\{\text{Ab}\}\|_{s=s_*}$	$P\{\text{Ab}\}\|_{s=s_\bullet}$
10	10.12	0.11	10.36	0.02	10.38	1.874×10^{-1}	1.863×10^{-1}
20	18.38	-0.31	18.63	-0.05	18.58	1.920×10^{-1}	1.937×10^{-1}
50	42.57	-1.01	42.83	-0.13	42.69	1.961×10^{-1}	1.981×10^{-1}
100	82.53	-1.72	82.78	-0.17	82.61	1.978×10^{-1}	1.993×10^{-1}
200	162.30	-2.65	162.54	-0.19	162.35	1.989×10^{-1}	1.998×10^{-1}
500	401.70	-4.39	401.91	-0.18	401.73	1.996×10^{-1}	2.000×10^{-1}
1000	801.12	-6.28	801.28	-0.15	801.13	1.998×10^{-1}	2.000×10^{-1}

3.6 Conclusions

The analytical assessment and numerical experiments in Sections 3.3 and 3.4 clearly suggest that the first-order diffusion approximations and conventional square-root staffing with respect to the tail probability of the customer delay are less accurate for overloaded systems. It is shown that significant $\beta_\bullet$ values arise when $\beta_* < 0$ (especially when β_* is relatively small or more negative), while $\beta_* > 0$ is typically associated with a small $\beta_\bullet$. In these two

types of constraint satisfaction problems, $\beta_* < 0$ can be due to different system parameters, such as a large ϵ (i.e., a loose constraint), a large λ (due to economy of scale), and/or a large θ (more "contribution" from customer abandonment). In these cases, the refinement term (in either the approximation or staffing) significantly improves the accuracy, and such an improvement leads to the right staffing level in most cases of practical interest to call center staffing.

Although ED+QED staffing is more accurate than conventional square-root staffing when the constraint satisfaction problem leads to an overloaded system, refined square-root staffing is very accurate in all cases (in particular, about as accurate as ED+QED in the overloaded case) and thus overall the most reliable method, at least under our model assumptions.

As for staffing under the abandonment constraint, we observe in Section 3.5 that the refinement can be significant when the constraint is tight, regardless of the customer patience level or the system size. In all our experiments, the refined square-root staffing rule yields satisfactory results.

3.7 Discussion

As discussed in Subsection 3.2.1 and demonstrated throughout this chapter, the refined square-root staffing approach is based upon performance approximation results of the same form as (145). In principle, this approach can be applied to optimization problems that are much more complicated than those considered above, as long as one can develop the relevant performance measure approximation results with the same form as (145). For example, in [30], Janssen et al. also consider a cost optimization problem and they essentially reduce that problem to a constraint satisfaction problem, in which the constraint to meet is that the derivative of the objective function in the optimization problem with respect to the decision variable equals zero.

One potential challenge in implementing the refined square-root staffing approach to solving a more complex problem, however, is that the expression of the refinement term for the optimization solution can be very complicated and therefore its calculation may

not be as straightforward as we have seen so far. In what follows, we briefly describe a capacity-inventory joint optimization problem, where such complication arises, and discuss a heuristic for addressing this challenge.

3.7.1 Managing Capacity and Inventory Jointly for Large Manufacturing Systems

Consider a single-product, make-to-stock manufacturing system with the production facility modeled as a parallel-server queue. Inventory is managed under a base-stock policy with base stock level s and the production facility consists of c servers in parallel. Customer orders arrive according to a Poisson process with rate λ and are fulfilled from the finished-goods inventory if the product is available. If no on-hand inventory is available at the time of an order arrival, the order is backlogged and a new production order is placed to the facility. Production orders are processed at the facility (i.e., c-server queue) on a first-come-first-serve basis. The amount of time to process each order is exponentially distributed with rate μ and they are independent of one another. We assume $\mu = 1$ without loss of generality. Denote by h the unit inventory holding cost per unit time, by p the unit backorder penalty cost per unit time, by w the unit work-in-process (WIP) cost per unit time at the production facility, and by d the cost per unit time for running one server.

Our objective is to determine the production capacity size c and the inventory base stock level s in order to minimize the total long-run average cost per unit time in the system. Specifically, we want to choose a pair of non-negative integers (c, s) so as to minimize the cost objective function

$$\Pi(c, s, R) := d \cdot c + w \cdot \mathrm{E}[Q_c] + h \cdot \mathrm{E}[(s - Q_c)^+] + p \cdot \mathrm{E}[(Q_c - s)^+], \tag{190}$$

where Q_c is simply equal in distribution to the steady-state number of customers in the M/M/c queue. We remark that this problem is a parallel-server version of the one considered in [9]. To simplify the exposition, in the remainder of this section we shall assume that c and s can take on non-integer values and do not distinguish between each performance measure function and its analytic continuation, which all can be properly defined (see [44]).

We first rewrite

$$\Pi(c,s,\lambda) = (d+w)\lambda + \lambda^{1/2}K(c,s,\lambda), \tag{191}$$

where

$$K(c,s,\lambda) := d\cdot\frac{c-\lambda}{\sqrt{\lambda}} + w\cdot\frac{\mathrm{E}[Q_c]-\lambda}{\sqrt{\lambda}} + h\cdot\frac{\mathrm{E}[(s-Q_c)^+]}{\sqrt{\lambda}} + p\cdot\frac{\mathrm{E}[(Q_c-s)^+]}{\sqrt{\lambda}}, \tag{192}$$

and focus on the objective function $K(c,s,\lambda)$ instead, since the first term in (191), $(d+w)\lambda$, is independent of c or s. The minimizer of $K(c,s,\lambda)$, say $(c_{\mathrm{opt}}, s_{\mathrm{opt}})$, must satisfy

$$\left(\frac{\partial K(c_{\mathrm{opt}},s_{\mathrm{opt}},\lambda)}{\partial c}, \frac{\partial K(c_{\mathrm{opt}},s_{\mathrm{opt}},\lambda)}{\partial s}\right) = (0,0), \tag{193}$$

assuming that it is a local optimal point. Then, in order to apply the refined square-root staffing approach, one would first derive an approximation result with the same form as (145), namely, for any $\beta > 0$ and $b \in (-\infty,\infty)$,

$$K(\lambda+\beta\sqrt{\lambda},\lambda+b\sqrt{\lambda},\lambda) = K_*(\beta,b) + K_\bullet(\beta,b)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{194}$$

$$\frac{\partial K(\lambda+\beta\sqrt{\lambda},\lambda+b\sqrt{\lambda},\lambda)}{\partial\beta} = \frac{\partial K_*(\beta,b)}{\partial\beta} + \frac{\partial K_\bullet(\beta,b)}{\partial\beta}\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{195}$$

and

$$\frac{\partial K(\lambda+\beta\sqrt{\lambda},\lambda+b\sqrt{\lambda},\lambda)}{\partial b} = \frac{\partial K_*(\beta,b)}{\partial b} + \frac{\partial K_\bullet(\beta,b)}{\partial b}\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{196}$$

where for any $\beta > 0$ and $b \in (-\infty,\infty)$,

$$K_*(\beta,b) := \begin{cases} d\beta - pb + \frac{w+p}{\beta}\cdot C_*(\beta) + (p+h)D_*(\beta,b)\left[b + \frac{\phi(b)}{\Phi(b)}\right], \text{if } b \le \beta \\ d\beta + hb + \frac{w-h}{\beta}\cdot C_*(\beta) + (p+h)[1-C_*(\beta)]\frac{\phi(\beta)}{\beta^2\Phi(\beta)}\cdot e^{-\beta(b-\beta)}, \text{if } b > \beta \end{cases} \tag{197}$$

$$D_*(\beta,b) := \begin{cases} [1-C_*(\beta)]\Phi(b)/\Phi(\beta) = \beta\Phi(b)[\phi(\beta)+\beta\Phi(\beta)]^{-1}, & \text{if } b \le \beta \\ 1 - C_*(\beta)e^{-\beta(b-\beta)}, & \text{if } b > \beta \end{cases}, \tag{198}$$

$$C_*(\beta) := \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}. \tag{199}$$

For simplicity we omit the expression of the $K_\bullet(\cdot,\cdot)$ function and refer to [44] for such details as well as the derivation of the above expansions. Based on these series expansion results, a two-dimensional analogue of expression (163) and similar optimality results are further

derived in [44] by following the same procedure as in the proof of Theorem 3.3.3 in two dimensions.

Specifically, the first step is to obtain a $\mathcal{O}(1)$-approximation for $(c_{\text{opt}}, s_{\text{opt}})$, namely (c_*, s_*). We recommend the same search procedure as followed in solving the capacity-inventory joint optimization problem for a single-server facility by Bradley and Glynn in [9]. First, it follows from the characterization of the newsvendor problem solution that $\frac{\partial K(c,s,\lambda)}{\partial s} = 0$ is equivalent to $\mathrm{P}\{Q_c \leq s\} = p/(p+h)$. Also, $\frac{\partial K_*(\beta,b)}{\partial b} = 0$ is equivalent to

$$D_*(\beta, b) = \frac{p}{p+h}, \tag{200}$$

and the $D_*(\cdot,\cdot)$ function satisfies (see Proposition 2 in [24]) that, for any $\beta > 0$ and $b \in (-\infty, \infty)$,

$$\mathrm{P}\{Q_{\lambda+\beta\sqrt{\lambda}} \leq \lambda + b\sqrt{\lambda}\} = D_*(\beta, b) + \mathcal{O}(1). \tag{201}$$

If we define

$$w_*(\beta) := \begin{cases} \beta + \frac{1}{\beta} \ln\left[\frac{C_*(\beta)\cdot(p+h)}{h}\right], & \text{if } C_*(\beta) > \frac{h}{p+h} \\ \Phi^{-1}\left(\frac{p\Phi(\beta)}{(p+h)[1-C_*(\beta)]}\right), & \text{if } C_*(\beta) \leq \frac{h}{p+h} \end{cases}, \tag{202}$$

it is easy to verify that $D_*(\beta, w_*(\beta)) = p/(p+h)$ or equivalently

$$w_*(\beta) = \arg \min_{b\in(-\infty,\infty)} K_*(\beta, b). \tag{203}$$

Hence, we recommend numerically searching for

$$\beta_* := \arg\min_{\beta>0} K_*(\beta, w_*(\beta)), \tag{204}$$

calculating $b_* = w_*(\beta_*)$ and then setting $(c_*, s_*) = (\lambda + \beta_*\sqrt{\lambda}, \lambda + b_*\sqrt{\lambda})$. Note that $K_*(\beta, w_*(\beta))$ in general is not convex in β and one can only obtain β_* numerically.

The second step is to obtain the refinement, say, $(\beta_\bullet, b_\bullet)$, and then the refined approximative optimal solution $(c_\bullet, s_\bullet) := (c_*, s_*) + (\beta_\bullet, b_\bullet)$. It is shown in [44] that a two-dimensional analogue of expression (163) reads

$$(\beta_\bullet, b_\bullet) := J^{-1} \cdot \left(\frac{\partial K_\bullet(\beta_*, b_*)}{\partial b} \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta \partial b} - \frac{\partial^2 K_*(\beta_*, b_*)}{\partial b^2} \frac{\partial K_\bullet(\beta_*, b_*)}{\partial \beta}, \right.$$
$$\left. \frac{\partial^2 K_*(\beta_*, b_*)}{\partial b \partial \beta} \frac{\partial K_\bullet(\beta_*, b_*)}{\partial \beta} - \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta^2} \frac{\partial K_\bullet(\beta_*, b_*)}{\partial b} \right), \tag{205}$$

where

$$J := \frac{\partial^2 K_*(\beta_*, b_*)}{\partial b^2} \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta^2} - \frac{\partial^2 K_*(\beta_*, b_*)}{\partial b \partial \beta} \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta \partial b}. \tag{206}$$

However, we find that expression (205) is extremely complicated because it involves first-order and second-order derivatives of the $K_*(\cdot,\cdot)$ and $K_\bullet(\cdot,\cdot)$ functions, both of which themselves already have rather complex expressions. As a consequence, standard computational software packages such as Mathematica have difficulties evaluating (205), although the refinement prescription (205) has the advantage of being independent of λ and hence needs to be calculated only once for problem instances with different demand rates. We are currently investigating this issue.

3.7.2 A Heuristic

In this subsection, we discuss a simple heuristic that in some situations may help to resolve the challenge encountered in the previous subsection, i.e., the calculation of the refinement term is too complicated. To illustrate the main idea, we shall present the heuristic in the context of the delay constraint problem for the Erlang A model.

Suppose the goal is to solve the delay constraint problem for a large value of λ (say, at the order of 100's or 1000's). Further suppose that (154) were too complex to evaluate and thus one could not calculate $s_\bullet$ directly. In this case, one may consider the following procedure, if the exact optimum for a medium-size problem (i.e., λ at the order of 10's) can be obtained relatively easily.

First, fix a medium value $\lambda_m > 0$, and solve for s_m such that $A(s_m, \lambda_m, \theta) = \epsilon$. Next, solve for β_* such that $A_*(\beta_*) = \epsilon$. Then for any large λ, to meet the constraint $A(s, \lambda, \theta) \leq \epsilon$, recommend the staffing level $\lceil \hat{s}_\bullet \rceil$, where

$$\hat{s}_\bullet := s_* + \hat{\beta}_\bullet, \tag{207}$$

with

$$s_* := \lambda + \beta_* \sqrt{\lambda} \quad \text{and} \quad \hat{\beta}_\bullet := s_m - (\lambda_m + \beta_* \sqrt{\lambda_m}). \tag{208}$$

As suggested by the notation that we use, the main idea here is to use $\hat{\beta}_\bullet$ as an estimator for $\beta_\bullet$.

Define $f(\lambda) := s_{\text{opt}} - s_\bullet$. Then the optimality gap of $\hat{s}_\bullet$ has the following characterization.

Proposition 3.7.1.

$$s_{\text{opt}} - \hat{s}_\bullet = f(\lambda) - f(\lambda_m) = -f(\lambda_m) + \mathcal{O}(\lambda^{-1/2}) = \mathcal{O}(1). \tag{209}$$

Proof. It follows from Theorem 3.3.3 that $f(\lambda) = \mathcal{O}(\lambda^{-1/2})$. Also, we have

$$\hat{\beta}_\bullet = s_m - (\lambda_m + \beta_* \sqrt{\lambda_m}) = \beta_\bullet + f(\lambda_m). \tag{210}$$

Hence,

$$s_{\text{opt}} - \hat{s}_\bullet = [s_* + \beta_\bullet + f(\lambda)] - [s_* + \beta_\bullet + f(\lambda_m)] = f(\lambda) - f(\lambda_m) = -f(\lambda_m) + \mathcal{O}(\lambda^{-1/2}). \tag{211}$$

□

Recall that

$$s_{\text{opt}} - s_* = \beta_\bullet + f(\lambda) = \beta_\bullet + \mathcal{O}(\lambda^{-1/2}) = \mathcal{O}(1). \tag{212}$$

Comparing (209) with (212), we find that both $\hat{s}_\bullet$ and s_* are optimal up to $\mathcal{O}(1)$ and therefore neither of them is necessarily more accurate than the other. More specifically, while the $\beta_\bullet$ term is captured by $\hat{\beta}_\bullet$, an error of $f(\lambda_m)$ is introduced into $\hat{s}_\bullet$. However, if λ_m is chosen to be sufficiently large, this $f(\lambda_m)$ error is negligible, as $f(\lambda)$ converges to 0 at the rate of $\lambda^{-1/2}$; on the other hand, the larger λ_m, the more expensive is solving for s_m and obtaining $\hat{\beta}_\bullet$ computationally. Our numerical experiments suggest that choosing λ_m at around 20 to 40 seems to work very well; see Table 16.

Table 16: $A(s, \lambda, \theta) = 0.5$, with $\theta = 80$; $\lambda_m = 25$ is used to calculate $\hat{\beta}_\bullet$.

λ	s_{opt}	β_*	s_*	$s_{\text{opt}} - s_*$	$\hat{\beta}_\bullet$	$\hat{s}_\bullet$	$s_{\text{opt}} - \hat{s}_\bullet$
25	15.27	-4.38	3.05	12.22	12.22	15.27	0.00
125	89.33	-4.38	75.93	13.39	12.22	88.15	1.17
225	172.48	-4.38	159.17	13.30	12.22	171.39	1.08
325	259.12	-4.38	245.88	13.23	12.22	258.10	1.01
425	347.71	-4.38	334.52	13.18	12.22	346.75	0.96
525	437.58	-4.38	424.44	13.14	12.22	436.66	0.92
625	528.40	-4.38	515.28	13.11	12.22	527.50	0.89
725	619.92	-4.38	606.83	13.08	12.22	619.05	0.86
825	712.01	-4.38	698.95	13.06	12.22	711.17	0.84
925	804.58	-4.38	791.53	13.04	12.22	803.75	0.82
1025	897.53	-4.38	884.50	13.03	12.22	896.72	0.81

In summary, we have proposed that one attack a problem instance with large λ by solving the diffusion approximation counterpart of the problem (which is independent of λ) and, in addition, obtaining the exact optimum for another problem instance with a moderate value of λ. The diffusion approximation part of this procedure yields s_*, which not only serves as a $\mathcal{O}(1)$ approximation for the true optimum but acts as a reference point to compare the medium-λ problem solution with in order to obtain an estimator of $\beta_\bullet$. Although not guaranteed to yield a solution more accurate than s_* itself, this heuristic may be effective in some cases, in particular, if (1) the refinement term turns out to be significant, and (2) solving a medium-λ problem is relatively easy computationally while solving a large-λ one becomes formidably difficult.

3.8 Proofs

This section contains the proofs of most results presented in this chapter.

3.8.1 Proofs for the delay constraint problem

We first establish a lemma which will be used several times in the proof. In the statement of Lemma 3.8.1, a function $g_1(x) = \mathcal{O}(g_2(x))$, if $\limsup_{x\to 0} |g_1(x)/g_2(x)| < \infty$.

Lemma 3.8.1. *If*

$$f(x) = f_1(x) + f_2(x) \cdot x + \mathcal{O}(x^2), \tag{213}$$

where both $f_1(x)$ and $f_2(x)$ are $\mathcal{O}(1)$, then $f(x)^{-1} = f_1(x)^{-1} - f_2(x) \cdot f_1(x)^{-2}x + \mathcal{O}(x^2)$.

Proof. Define $k(x) := f_1(x)^{-1} - f_2(x) \cdot f_1(x)^{-2}x$. We then have $f(x) \cdot k(x) = 1 + \mathcal{O}(x^2)$ and thus

$$f(x) \cdot [f(x)^{-1} - k(x)] = 1 - f(x) \cdot k(x) = \mathcal{O}(x^2). \tag{214}$$

Combining (214) with $f(x) = \mathcal{O}(1)$ yields the desired result. □

Next, we prove Theorem 3.3.1.

Proof of Theorem 3.3.1. Throughout this proof, let $s = \lambda + \beta\sqrt{\lambda}$ and thus $s = \mathcal{O}(\lambda)$. We denote the upper incomplete gamma function by

$$\Gamma(s, a) = \int_a^\infty t^{s-1}e^{-t}dt, \tag{215}$$

and the gamma function by $\Gamma(s) = \gamma(s, a) + \Gamma(s, a)$. Using the relation

$$\gamma(s, \lambda) = \frac{\lambda^s e^{-\lambda}}{s} + \Gamma(s)\left(1 - \frac{\Gamma(s+1, \lambda)}{\Gamma(s+1)}\right) \tag{216}$$

and (143) yields

$$\frac{se^\lambda}{\lambda^s}\gamma(s, \lambda) = 1 + \frac{\Gamma(s+1)e^\lambda}{\lambda^s} - B(s, \lambda)^{-1}. \tag{217}$$

First, we have that

$$B(s, \lambda)^{-1} = \frac{\Phi(\alpha)}{\phi(\alpha)}s^{1/2} + \frac{2}{3} + \mathcal{O}(s^{-1/2}), \tag{218}$$

where

$$\alpha = \sqrt{-2s(1 - \rho + \ln \rho)}, \quad \text{sign}(\alpha) = \text{sign}(1 - \rho), \quad \text{and} \quad \rho = \lambda/s. \tag{219}$$

For any positive integer s, (218) holds due to Theorem 1 in [29] and the fact that $\alpha \to \beta$ as $\lambda \to \infty$; for any real $s > \lambda$, (218) follows from Theorem 1 and relation (6.1) in [30]. By letting $p(s) := s^s e^{-s}\sqrt{2\pi s}\ \Gamma(s+1)^{-1}$, we rewrite the second term in (217) as

$$\frac{\Gamma(s+1)e^{\lambda}}{\lambda^s} = \frac{s^{1/2}}{\phi(\alpha)p(s)}. \tag{220}$$

Stirling's formula for the gamma function (see page 257 of [1]) reads

$$\Gamma(s+1) = s\Gamma(s) = s^s e^{-s}\sqrt{2\pi s}[1 + \mathcal{O}(s^{-1})], \tag{221}$$

and thus

$$p(s)^{-1} = 1 + \mathcal{O}(s^{-1}). \tag{222}$$

Applying (222), (218), (220), and $\phi(\alpha)^{-1} = \mathcal{O}(1)$ to (217) yields

$$\frac{se^{\lambda}}{\lambda^s}\gamma(s,\lambda) = \frac{\Phi(-\alpha)}{\phi(\alpha)}s^{1/2} + \frac{1}{3} + \mathcal{O}(s^{-1/2}). \tag{223}$$

We then multiply both sides of (223) by $s^{-1/2}$ to obtain

$$\frac{se^{\lambda}}{\lambda^s}\gamma(s,\lambda)\cdot s^{-1/2} = \frac{\Phi(-\alpha)}{\phi(\alpha)} + \frac{1}{3}s^{-1/2} + \mathcal{O}(s^{-1}). \tag{224}$$

As noted above, as $s \to \infty$ (or equivalently $\lambda \to \infty$), α converges to β, where β does not change with s (or λ). Therefore, $\Phi(-\alpha)/\phi(\alpha) = \mathcal{O}(1)$. This allows us to apply Lemma 3.8.1 in taking the reciprocal of (224) to arrive at

$$\frac{\lambda^s e^{-\lambda}}{s\gamma(s,\lambda)}\cdot s^{1/2} = \frac{\phi(\alpha)}{\Phi(-\alpha)} - \frac{1}{3}\left(\frac{\phi(\alpha)}{\Phi(-\alpha)}\right)^2 s^{-1/2} + \mathcal{O}(s^{-1}), \tag{225}$$

and therefore

$$\frac{\lambda^s e^{-\lambda}}{s\gamma(s,\lambda)} = \frac{\phi(\alpha)}{\Phi(-\alpha)}s^{-1/2} - \frac{1}{3}\left(\frac{\phi(\alpha)}{\Phi(-\alpha)}\right)^2 s^{-1} + \mathcal{O}(s^{-3/2}). \tag{226}$$

Substituting (226) and (218) into (141) (with s replaced by $\lambda + \beta\sqrt{\lambda}$) then yields

$$A(\lambda + \beta\sqrt{\lambda}, \lambda, \theta)^{-1} = A_*(\alpha)^{-1}\left(1 - \frac{1}{3}\sqrt{\theta}H_\theta(\alpha)s^{-1/2}\right) + \mathcal{O}(s^{-1}). \tag{227}$$

Let us recall the definition $G(x) = \Phi(x)/\phi(x)$, for any $x \in (-\infty, \infty)$. Then simple computations show that

$$G(\alpha) = G(\beta) - \frac{1}{6}\beta^2\left(1 + \beta G(\beta)\right)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{228}$$

$$\phi(\alpha)^{-1} = \phi(\beta)^{-1} - \frac{1}{6}\beta^3\phi(\beta)^{-1}\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{229}$$

and $s^{-1/2} = \lambda^{-1/2} + \mathcal{O}(\lambda^{-1})$ due to the relation $s = \lambda + \beta\sqrt{\lambda}$, where β does not scale with λ. Subtracting (228) from (229) yields

$$\frac{\Phi(-\alpha)}{\phi(\alpha)} = \frac{\Phi(-\beta)}{\phi(\beta)} + \frac{1}{6}\beta^2 \left(1 - \frac{\beta\Phi(-\beta)}{\phi(\beta)}\right) \lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{230}$$

Taking the reciprocal of (230) gives

$$\frac{\phi(\alpha)}{\Phi(-\alpha)} = \frac{\phi(\beta)}{\Phi(-\beta)} - \frac{1}{6}\beta^2 \left(\frac{\phi(\beta)^2}{\Phi(-\beta)^2} - \frac{\beta\phi(\beta)}{\Phi(-\beta)}\right) \lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{231}$$

Using (228) and (231) in (227), we arrive at

$$A_*(\alpha)^{-1} = A_*(\beta)^{-1} + h_\theta(\beta)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}) \tag{232}$$

and

$$1 - \frac{1}{3}\sqrt{\theta}H_\theta(\alpha)s^{-1/2} = 1 - \frac{1}{3}\sqrt{\theta}H_\theta(\beta)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{233}$$

Therefore, by multiplying (232) and (233), we obtain that

$$A(\lambda + \beta\sqrt{\lambda}, \lambda, \theta)^{-1} = A_*(\beta)^{-1} + \left(h_\theta(\beta) - \frac{1}{3}\sqrt{\theta}H_\theta(\beta)A_*(\beta)^{-1}\right) \lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{234}$$

Finally, taking the reciprocal of (234) yields (150). □

We next provide the proof of Proposition 3.3.2.

Proof of Proposition 3.3.2. First, $G(\beta)$ and $H_\theta(\beta)$ are always positive, and therefore by the definition $A_\bullet(\beta) = A_*(\beta)^2 \left(\frac{1}{3}\sqrt{\theta}H_\theta(\beta)A_*(\beta)^{-1} - h_\theta(\beta)\right)$, it suffices to prove $h_\theta(\beta) \le 0$. Recall that

$$h_\theta(\beta) = -\frac{1}{6}\sqrt{\theta}\beta^2 H_\theta(\beta) \left(G(\beta)H_\theta(\beta)\theta^{-1/2} - \beta G(\beta)\theta^{-1} + 1 + \beta G(\beta)\right) \tag{235}$$

$$= -\frac{1}{6}\sqrt{\theta}\beta^2 H_\theta(\beta)\Big[G(\beta)\theta^{-1/2} \cdot \Big[H_\theta(\beta) - \beta\theta^{-1/2}\Big] + [1 + \beta G(\beta)]\Big]. \tag{236}$$

By the property of the normal hazard rate function (see Section 5 of the Internet supplement to [62]),

$$H_\theta(\beta) - \beta\theta^{-1/2} \ge 0, \quad \text{for any } \theta > 0 \text{ and } \beta \in (-\infty, \infty). \tag{237}$$

In addition, for any $\beta \in (-\infty, \infty)$, we have that

$$1 + \beta G(\beta) \ge 0. \tag{238}$$

For $\beta \geq 0$, (238) obviously holds since both β and $G(\beta)$ are non-negative. For $\beta < 0$, due to the relation that $\Phi(-x) \leq \phi(x)/x$ for any $x > 0$ (see Section 5 of the Internet supplement to [62]), we have that $\Phi(\beta) \leq \phi(\beta)/(-\beta)$ and therefore (238) holds. Finally, substituting (237) and (238) into (236) yields $h_\theta(\beta) \leq 0$ and therefore $A_\bullet(\beta) > 0$.

Next, we prove $\lim_{\beta\to\infty} A_\bullet(\beta) = 0$. Applying L'Hopital's rule, we have

$$\lim_{\beta\to\infty} H_\theta(\beta)/(\beta\theta^{-1/2}) = 1 \tag{239}$$

or

$$H_\theta(\beta) = \beta\theta^{-1/2} \cdot [1 + o(1)], \quad \text{as } \beta \to \infty, \tag{240}$$

where the $o(1)$ terms in (240) and in the next few expressions in this proof all denote a term that converges to zero as $\beta \to \infty$. Similarly, we obtain that

$$G(\beta) = \frac{1}{\phi(\beta)} - \frac{\Phi(-\beta)}{\phi(\beta)} = \frac{1}{\phi(\beta)} - \frac{1}{\beta} \cdot [1 + o(1)], \quad \text{as } \beta \to \infty. \tag{241}$$

Applying (240) and (241) to (236), we have that

$$h_\theta(\beta) = -\frac{1}{6}\beta^3 \cdot \frac{\beta}{\phi(\beta)} \cdot [1 + o(1)], \quad \text{as } \beta \to \infty. \tag{242}$$

It is known that $\lim_{\beta\to\infty} A_*(\beta) = 0$ (see the e-companion of [40]). In fact, substituting (240) and (241) into the definition $A_*(\beta) = [1 + \sqrt{\theta}G(\beta)H_\theta(\beta)]^{-1}$ yields

$$A_*(\beta) = \frac{\phi(\beta)}{\beta} \cdot [1 + o(1)], \quad \text{as } \beta \to \infty. \tag{243}$$

Finally, $\lim_{\beta\to\infty} A_\bullet(\beta) = 0$ follows from applying (240), (242), (243), and $\lim_{\beta\to\infty} \phi(\beta)\beta^m = 0$, for any positive integer m, to the definition of $A_\bullet(\beta)$ (152).

To prove the last assertion $\lim_{\beta\to-\infty} A_\bullet(\beta) = 0$, we first obtain by simple calculations that for any integer $m \geq 0$,

$$\lim_{\beta\to-\infty} \beta^m H_\theta(\beta) = \lim_{\beta\to-\infty} G(\beta) = 0 \quad \text{and} \quad \lim_{\beta\to-\infty} \beta G(\beta) = -1. \tag{244}$$

Applying (244) to (152) immediately yields the desired result. □

3.8.2 Proofs for excess delay constraint

We first show a technical lemma, which is needed in the later proof.

Lemma 3.8.2. *Let*

$$v_\lambda(x) = \exp\{-b_1\sqrt{\lambda}x - b_2\lambda x^2\}(1 + b_3\lambda x^3), \tag{245}$$

$$w_\lambda(x) = \exp\{-b_1\sqrt{\lambda}x - b_2\lambda x^2 + b_3\lambda x^3\}, \tag{246}$$

where $b_i > 0$, $i = 1, 2, 3$, are constants. Let $t \geq 0$ and $\delta \in (t\lambda^{-1/2}, b_2/b_3)$ be a constant, and define

$$I(\lambda) = \int_{t\lambda^{-1/2}}^{\delta} w_\lambda(x)dx, \qquad I_A(\lambda) = \int_{t\lambda^{-1/2}}^{\infty} v_\lambda(x)dx. \tag{247}$$

Then,

$$\begin{aligned} I(\lambda) &= I_A(\lambda) + \mathcal{O}(\lambda^{-3/2}), & (248)\\ I_A(\lambda) &= \frac{\Phi\Big(-\sqrt{2b_2}\ t - \frac{1}{\sqrt{2}}b_1 b_2^{-1/2}\Big)}{\phi\Big(\frac{1}{\sqrt{2}}b_1 b_2^{-1/2}\Big)\sqrt{2b_2}}\lambda^{-1/2} + I_\bullet(b_1, b_2, t)b_3\lambda^{-1}, & (249) \end{aligned}$$

where, $\forall a > 0, b > 0, t \geq 0$,

$$\begin{aligned} I_\bullet(a, b, t) &= \int_t^\infty \exp\{-ay - by^2\}y^3 dy \\ &= \frac{1}{16}b^{-7/2}e^{-t(a+bt)}\Big[2\sqrt{b}\left(a^2 - 2abt + 4b\left(1 + bt^2\right)\right) \\ &\quad + a\left(a^2 + 6b\right)e^{(a+2bt)^2/4b}\sqrt{\pi}\Big(\mathrm{Erf}\Big[\frac{1}{2}b^{-1/2}(a + 2bt)\Big] - 1\Big)\Big], \end{aligned} \tag{250}$$

with $\mathrm{Erf}(x) := 2\pi^{-1/2}\int_0^x \exp\{-t^2/2\}dt$.

Proof. We have that

$$\begin{aligned} I(\lambda) &= \int_{t\lambda^{-1/2}}^{\delta} w_\lambda(x)dx \\ &= \lambda^{-1/2}\int_t^{\delta\sqrt{\lambda}} \exp\{-b_1 y - b_2 y^2 + b_3 y^3\lambda^{-1/2}\}dy \\ &= z\int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2 + b_3 y^3 z\}dy \end{aligned} \tag{251}$$

with $y = x\sqrt{\lambda}$ and $z = \lambda^{-1/2}$. By Taylor series expansion, for some $\xi \in (0, z)$,

$$\exp\{-b_1 y - b_2 y^2 + b_3 y^3 z\} = \exp\{-b_1 y - b_2 y^2\}\Big(1 + b_3 y^3 z + \frac{1}{2}b_3^2 y^6 e^{b_3 y^3 \xi}z^2\Big).$$

Therefore,

$$I(\lambda) = I_1(\lambda) + I_2(\lambda), \tag{252}$$

where

$$I_1(\lambda) = z\int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2\}(1 + b_3 y^3 z)dy, \tag{253}$$

$$I_2(\lambda) = \frac{1}{2}z^3 \int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2\} b_3^2 y^6 e^{b_3 y^3 \xi} dy. \tag{254}$$

Let $I_{A_1}(\lambda) = \int_{t\lambda^{-1/2}}^{\delta} v_\lambda(x)dx$ and $I_{A_2}(\lambda) = \int_\delta^\infty v_\lambda(x)dx$, and then we have

$$I_A(\lambda) = I_{A_1}(\lambda) + I_{A_2}(\lambda). \tag{255}$$

Upon a change of variables, we have that

$$I_1(\lambda) = I_{A_1}(\lambda). \tag{256}$$

The fact that

$$I_2(\lambda) = \mathcal{O}(\lambda^{-3/2}) \tag{257}$$

follows from

$$\begin{aligned}
I_2(\lambda) &= \frac{1}{2}z^3 \int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2\} b_3^2 y^6 e^{b_3 y^3 \xi} dy \\
&\le \frac{1}{2}z^3 \int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2\} b_3^2 y^6 e^{b_3 y^2 \frac{\delta}{z} z} dy \quad \text{(because } y \le \delta/z \text{ and } \xi \le z\text{)} \\
&= \frac{1}{2}z^3 \int_t^{\delta/z} \exp\{-b_1 y - b_2 y^2 + b_3 y^2 \delta\} b_3^2 y^6 dy \\
&\le \frac{1}{2}z^3 \int_0^{\infty} \exp\{-b_1 y - (b_2 - b_3\delta) y^2\} b_3^2 y^6 dy \\
&= C_0 z^3 = C_0 \lambda^{-3/2},
\end{aligned} \tag{258}$$

for some constant $C_0 > 0$, because we assume $\delta < b_2/b_3$ or $b_2 - b_3\delta > 0$.

Next we show that

$$I_{A_2}(\lambda) = o(e^{-\lambda^{\nu_0}}), \quad \text{for some } \nu_0 > 0. \tag{259}$$

For an arbitrarily chosen $C_1 \in (0,1)$, $\exists \lambda_{b_1,b_2,b_3,C_1} > 0$ such that, for any $\lambda > \lambda_{b_1,b_2,b_3,C_1}$, $v_\lambda(x) < \exp\{-b_2 C_1 \lambda x^2\}$. After integration, $I_{A_2}(\lambda) \le \int_\delta^\infty \exp\{-b_2 C_1 \lambda x^2\}dx$, for any $\lambda >$

$\lambda_{b_1,b_2,b_3,C_1}$. Then by Lemma 4.3 in the Internet supplement to [62], we have $\int_\delta^\infty \exp\{-b_2 C_1 \lambda x^2\} dx = o(e^{-\lambda^{\nu_0}})$, for some $\nu_0 > 0$, and thus (259) follows.

Using (256), (257), and (259), we subtract (255) from (252) and arrive at

$$I(\lambda) - I_A(\lambda) = \mathcal{O}(\lambda^{-3/2}) + o(e^{-\lambda^{\nu_0}}) = \mathcal{O}(\lambda^{-3/2}). \tag{260}$$

Expression (249) follows from straightforward calculations. Specifically,

$$\begin{aligned} I_A(\lambda) =& \int_{t\lambda^{-1/2}}^\infty \exp\{-b_1\sqrt{\lambda}x - b_2\lambda x^2\} dx + \int_{t\lambda^{-1/2}}^\infty \exp\{-b_1\sqrt{\lambda}x - b_2\lambda x^2\} b_3 \lambda x^3 dx \\ =& \lambda^{-1/2} \int_t^\infty \exp\{-b_1 y - b_2 y^2\} dy + \lambda^{-1/2} \int_t^\infty \exp\{-b_1 y - b_2 y^2\} b_3 y^3 \lambda^{-1/2} dy, \end{aligned} \tag{261}$$

where (261) is due to a change of variables $y = \sqrt{\lambda}x$ in both integral terms. Note that the second term in (261) is simply $I_\bullet(b_1, b_2, t) b_3 \lambda^{-1}$. We then apply another change of variables $z = \sqrt{2b_2} \cdot (y + b_1/2b_2)$ to the first term of (261) and obtain that

$$\begin{aligned} I_A(\lambda) &= (2b_2\lambda)^{-1/2} \cdot \exp\{b_1^2/4b_2\} \cdot \int_{\sqrt{2b_2}\ t + \frac{1}{\sqrt{2}} b_1 b_2^{-1/2}}^\infty \exp\{-z^2\} dz + I_\bullet(b_1, b_2, t) b_3 \lambda^{-1} \\ &= \frac{\Phi\Big(-\sqrt{2b_2}\ t - \frac{1}{\sqrt{2}} b_1 b_2^{-1/2}\Big)}{\phi\Big(\frac{1}{\sqrt{2}} b_1 b_2^{-1/2}\Big)\sqrt{2b_2}} \lambda^{-1/2} + I_\bullet(b_1, b_2, t) b_3 \lambda^{-1}. \end{aligned} \tag{262}$$

This establishes (249) and thus completes the proof of the lemma. □

Next, we briefly outline the proof of Theorem 3.4.1. We shall write $P\{W > t\lambda^{-1/2}\}$ as the product of $P\{W > 0\}$ and $P\{W > t\lambda^{-1/2} | W > 0\}$. Because we have obtained the asymptotic expansion for $P\{W > 0\}$ in Theorem 3.3.1, the main step in expanding $P\{W > t\lambda^{-1/2}\}$ is just to derive a refined approximation for the conditional probability $P\{W > t\lambda^{-1/2} | W > 0\}$. Define

$$u_\lambda(x) = \lambda\theta^{-1}(1 - e^{-\theta x}) - \lambda x - \beta\sqrt{\lambda}x, \tag{263}$$

$$J(y) = \int_y^\infty \exp\{u_\lambda(x)\} dx, \quad \forall\ y \geq 0. \tag{264}$$

From Equations (9.7) and (9.15) in [62], we have that, for $\forall\ t > 0$,

$$P\{W > t\lambda^{-1/2} | W > 0\} = \frac{e^{-\theta t \lambda^{-1/2}} J(t\lambda^{-1/2})}{J(0)}, \tag{265}$$

where $J(t\lambda^{-1/2})$ and $J(0)$ are key components of the conditional probability expression. Therefore, the first step of our proof is to obtain asymptotic expansions for $J(t\lambda^{-1/2})$ and $J(0)$. Then, we shall apply to (265) the approximations for $J(t\lambda^{-1/2})$ and $J(0)$ as well as a Taylor expansion of the term $e^{-\theta t\lambda^{-1/2}}$, and this will lead to a refined approximation for $P\{W > t\lambda^{-1/2}|W > 0\}$. Finally, we combine this result with Theorem 3.3.1 to reach the desired result.

Lemma 3.8.3.

$$J(t\lambda^{-1/2}) = \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + \frac{1}{6}I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\cdot\theta^2\lambda^{-1} + \mathcal{O}(\lambda^{-\frac{3}{2}}), \tag{266}$$

$$J(0) = H_\theta(\beta)^{-1}\theta^{-1/2}\lambda^{-1/2} + \frac{1}{6}I_\bullet\Big(\beta, \frac{1}{2}\theta, 0\Big)\cdot\theta^2\lambda^{-1} + \mathcal{O}(\lambda^{-\frac{3}{2}}). \tag{267}$$

Proof. We start from

$$e^{-\theta x} = 1 - \theta x + \frac{1}{2}\theta^2x^2 - \frac{1}{6}\theta^3x^3 + o(x^3), \quad \text{as } x \to 0. \tag{268}$$

Therefore, $\forall\epsilon > 0$, $\exists\delta(\epsilon) > 0$, such that, for any $x \in [0, \delta(\epsilon)]$

$$\frac{|e^{-\theta x} - (1 - \theta x + \frac{1}{2}\theta^2x^2 - \frac{1}{6}\theta^3x^3)|}{x^3} \le \epsilon. \tag{269}$$

Combining (269) with (263), we have that

$$-\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)\lambda x^3 \le u_\lambda(x) \le -\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)\lambda x^3. \tag{270}$$

In particular, we only consider those $\epsilon \in (0, \frac{1}{6}\theta^3)$ (so that the coefficient $\frac{1}{6}\theta^3 - \epsilon$ in the lower bound part of (270) is positive) and choose $\delta(\epsilon)$ such that

$$\delta(\epsilon) \in \left(0, \frac{1}{2}\theta^2\left(\frac{1}{6}\theta^3 + \epsilon\right)^{-1}\right). \tag{271}$$

Note that (271) guarantees the condition "$\delta < b_2/b_3$" in Lemma 3.8.2 is satisfied, where δ is replaced by $\delta(\epsilon)$, and b_2 and b_3 are replaced by the coefficients of the second and third terms in (270), namely $\frac{1}{2}\theta$ and $\frac{1}{\theta}(\frac{1}{6}\theta^3 \pm \epsilon)$, respectively. With fixed ϵ and $\delta(\epsilon)$, let $\lambda(\epsilon, \delta) = t^2/\delta(\epsilon)^2$. Then, for any $\lambda > \lambda(\epsilon, \delta)$, we have

$$t\lambda^{-1/2} < \delta(\epsilon), \tag{272}$$

and thus, by (270), we have that

$$\int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\Big\{ -\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)\lambda x^3\Big\}dx \le \int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\{u_\lambda(x)\}dx$$
$$\le \int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\Big\{ -\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)\lambda x^3\Big\}dx. \tag{273}$$

We next rewrite the integral definition of $J(t\lambda^{-1/2})$ as

$$J(t\lambda^{-1/2}) = \int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\{u_\lambda(x)\}dx + \int_{\delta(\epsilon)}^{\infty} \exp\{u_\lambda(x)\}dx. \tag{274}$$

Note that our definition of $u_\lambda(\cdot)$ is a special case of (11.5) on p. 32 of [62] by setting, in their notation, $\mu = 1$ and $\bar{G}(u) = e^{-\theta u}$. Therefore, it follows from (11.10) on p. 33 of [62] that the second term on the right-hand side of (274) is $o(e^{-\nu_1\lambda})$ for some $\nu_1 > 0$, or $J(t\lambda^{-1/2}) = \int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\{u_\lambda(x)\}dx + o(e^{-\nu_1\lambda})$. Substituting this into (273) yields

$$\int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\Big\{ -\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)\lambda x^3\Big\}dx + o(e^{-\nu_1\lambda}) \le J(t\lambda^{-1/2})$$
$$\le \int_{t\lambda^{-1/2}}^{\delta(\epsilon)} \exp\Big\{ -\beta\sqrt{\lambda}x - \frac{1}{2}\theta\lambda x^2 + \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)\lambda x^3\Big\}dx + o(e^{-\nu_1\lambda}). \tag{275}$$

Now, (271) and (272) allow us to apply Lemma 3.8.2 to (275) (with δ replaced by $\delta(\epsilon)$, b_1 by β, b_2 by $\frac{1}{2}\theta$, and b_3 by $\frac{1}{\theta}(\frac{1}{6}\theta^3 \pm \epsilon)$), and it follows that

$$\frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\cdot\frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}) \le J(t\lambda^{-1/2})$$
$$\le \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\cdot\frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}), \tag{276}$$

where the $o(e^{-\nu_1\lambda})$ term in (275) is dropped since if a function $f(\lambda) = o(e^{-\nu_1\lambda})$, it must hold that $f(\lambda) = \mathcal{O}(\lambda^{-3/2})$. From (276), we have that, for fixed $\epsilon > 0$, $\exists\lambda_2(\epsilon) > \lambda(\epsilon, \delta)$ such that for any $\lambda > \lambda_2(\epsilon)$,

$$\frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\cdot\frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)(1-\epsilon)\lambda^{-1} \le J(t\lambda^{-1/2})$$
$$\le \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\cdot\frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)(1+\epsilon)\lambda^{-1} \tag{277}$$

or

$$\begin{aligned} I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\cdot\frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)(1-\epsilon)\lambda^{-1} &\le J(t\lambda^{-1/2}) - \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} \\ &\le I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\cdot\frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)(1+\epsilon)\lambda^{-1}. \end{aligned} \tag{278}$$

Letting $\lambda \to \infty$, we have that

$$\begin{aligned} I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 - \epsilon\Big)(1-\epsilon) &\le \liminf_{\lambda\to\infty} \left(\lambda J(t\lambda^{-1/2}) - \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\sqrt{\lambda}\right) \\ &\le \limsup_{\lambda\to\infty} \left(\lambda J(t\lambda^{-1/2}) - \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\sqrt{\lambda}\right) \\ &\le I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big) \cdot \frac{1}{\theta}\Big(\frac{1}{6}\theta^3 + \epsilon\Big)(1+\epsilon). \end{aligned} \tag{279}$$

Letting $\epsilon \to 0$ yields

$$\lim_{\lambda\to\infty} \left(\lambda J(t\lambda^{-1/2}) - \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\sqrt{\lambda}\right) = \frac{1}{6} I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\theta^2. \tag{280}$$

This implies that

$$J(t\lambda^{-1/2}) = \frac{\Phi(-\sqrt{\theta}t - \beta\theta^{-1/2})}{\phi(\beta\theta^{-1/2})\sqrt{\theta}}\lambda^{-1/2} + \frac{1}{6} I_\bullet\Big(\beta, \frac{1}{2}\theta, t\Big)\theta^2\lambda^{-1} + o(\lambda^{-1}), \tag{281}$$

and then, from (276), we know that this $o(\lambda^{-1})$ is indeed $\mathcal{O}(\lambda^{-3/2})$. This yields the desired result (266), and (267) follows by letting $t = 0$. □

Finally, we complete the proof of Theorem 3.4.1.

Proof of Theorem 3.4.1. A straightforward Taylor series expansion yields

$$e^{-\theta t\lambda^{-1/2}} = 1 - \theta t\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}), \tag{282}$$

Substituting (282), (266), and (267) into (265), we obtain that

$$P\{W > t\lambda^{-1/2} | W > 0\} = d_*(\beta, t) + d_\bullet(\beta, t)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{283}$$

Multiplying (150) with (283) yields (172). □

Next, we provide the proof of Theorem 3.4.2.

Proof of Theorem 3.4.2. First, the existence of a unique β_* is known from Theorem 4.1 in [40]. If

$$A_*(\beta_*)d_\bullet(\beta_*, t) + A_\bullet(\beta_*)d_*(\beta_*, t) \neq 0, \tag{284}$$

we follow the same procedure as for Theorem 3.3.3, by replacing $A(s, \lambda, \theta)$ with $A_t(s, \lambda, \theta)$, $A_*(\cdot)$ with $A_*(\cdot)d_*(\cdot)$, and $A_\bullet(\cdot)$ with $A_*(\cdot)d_\bullet(\cdot) + A_\bullet(\cdot)d_*(\cdot)$. Note that (284) corresponds

to $A_\bullet(\beta_*) \neq 0$ in the proof of Theorem 3.3.3, which is needed to derive (161). We omit further details.

If $\beta_\bullet$=0 or

$$A_*(\beta_*)d_\bullet(\beta_*, t) + A_\bullet(\beta_*)d_*(\beta_*, t) = 0, \tag{285}$$

relation (169), i.e., $\beta_{\text{opt}} - \beta_\lambda = \mathcal{O}(\lambda^{-1})$, remains valid, where β_λ is a solution to

$$A_*(\beta_\lambda)d_*(\beta_\lambda, t) + \left[A_*(\beta_\lambda)d_\bullet(\beta_\lambda, t) + A_\bullet(\beta_\lambda)d_*(\beta_\lambda, t)\right]\lambda^{-1/2} = \epsilon. \tag{286}$$

From the definition of β_* and (285), we know that $\beta_\lambda = \beta_*$ actually solves (286) with the $\lambda^{-1/2}$ order term on the left-hand side of the equation being zero.

Substituting β_λ with β_* in (169) then yields that $\beta_{\text{opt}} - \beta_* = \mathcal{O}(\lambda^{-1})$, which in turn implies $s_{\text{opt}} - s_* = \mathcal{O}(\lambda^{-1/2})$, a sufficient condition for (178). Finally, (179) follows from $s_\bullet = s_*$ and $s_{\text{opt}} - s_* = \mathcal{O}(\lambda^{-1/2})$. □

3.8.3 Proofs for abandonment constraint

Proof of Theorem 3.5.1. Throughout this proof, let $s = \lambda + \beta\sqrt{\lambda}$ and thus $s = \mathcal{O}(\lambda)$. From equation (A.2) in [39], we have that

$$P\{\text{Ab}|W > 0\} = \left(\rho s\theta^{-1}e^{\lambda/\theta}(\lambda/\theta)^{-s/\theta}\gamma(s/\theta, \lambda/\theta)\right)^{-1} + 1 - \rho^{-1}. \tag{287}$$

Let us recall the definition $H_\theta(x) = \phi(x/\sqrt{\theta})/\Phi(x/\sqrt{\theta})$. Because θ is a positive constant not scaling with λ, we substitute λ in (226) with λ/θ, s with s/θ, and consequently α with $\alpha/\sqrt{\theta}$ (by plugging λ/θ and s/θ into (219)), and then have that

$$\frac{(\lambda/\theta)^{s/\theta}e^{-\lambda/\theta}}{s\theta^{-1}\gamma(s/\theta, \lambda/\theta)} = \sqrt{\theta}H_\theta(\alpha)s^{-1/2} - \frac{1}{3}H_\theta(\alpha)^2\theta s^{-1} + \mathcal{O}(s^{-3/2}). \tag{288}$$

Also, we note that

$$\rho^{-1} = \mathcal{O}(1) \quad \text{and} \quad 1 - \rho^{-1} = \mathcal{O}(\lambda^{-\frac{1}{2}}) = \mathcal{O}(s^{-\frac{1}{2}}). \tag{289}$$

Substituting (288) into (287), we obtain that

$$P\{\text{Ab}|W > 0\} = 1 - \rho^{-1} + \sqrt{\theta}H_\theta(\alpha)\rho^{-1}s^{-1/2} - \frac{1}{3}\rho^{-1}H_\theta(\alpha)^2\theta s^{-1} + \mathcal{O}(s^{-3/2}), \tag{290}$$

where the $\mathcal{O}(s^{-3/2})$ term in (290) comes from multiplying ρ^{-1} with the $\mathcal{O}(s^{-3/2})$ in (288) due to the first equation in (289). The fact that α converges to β then allows us to apply Lemma 3.8.1 in taking the reciprocal of (227), which yields that

$$P\{W > 0\} = A(s, \lambda, \theta) = A_*(\alpha) + \frac{1}{3}\sqrt{\theta}A_*(\alpha)H_\theta(\alpha)s^{-1/2} + \mathcal{O}(s^{-1}). \tag{291}$$

By noting (289), $A_*(\alpha) = \mathcal{O}(1)$, and $H_\theta(\alpha) = \mathcal{O}(1)$, we multiply (290) and (291) to arrive at

$$P\{\text{Ab}\} = A_*(\alpha)(1-\rho^{-1}) + \frac{1}{3}(2+\rho)A_*(\alpha)H_\theta(\alpha)\sqrt{\theta}\rho^{-1}s^{-1/2} + \mathcal{O}(s^{-3/2}). \tag{292}$$

We then just need to derive the series expansion of (292). Taking the reciprocal of (232) yields

$$A_*(\alpha) = A_*(\beta) - h_\theta(\beta)A_*(\beta)^2\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{293}$$

Then, using $1-\rho^{-1} = -\beta\lambda^{-1/2}$, we have that

$$A_*(\alpha)(1-\rho^{-1}) = -A_*(\beta)\beta\lambda^{-1/2} + h_\theta(\beta)A_*(\beta)^2\beta\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}). \tag{294}$$

To expand the second term of (292), we first note that

$$s^{-1/2} = \lambda^{-1/2} - \frac{1}{2}\beta\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}). \tag{295}$$

Then, similar to how we derive (288) from (226), by replacing λ with λ/θ, s with s/θ, α with $\alpha/\sqrt{\theta}$, and β with $\beta/\sqrt{\theta}$ in the expression (231), we get the following expansion

$$H_\theta(\alpha) = H_\theta(\beta) - \frac{1}{6}\beta^2\theta^{-1/2}\left(H_\theta(\beta)^2 - \beta H_\theta(\beta)\theta^{-1/2}\right)\lambda^{-1/2} + \mathcal{O}(\lambda^{-1}). \tag{296}$$

Combining (295) with (296) and (293), we obtain that

$$\begin{aligned}\frac{1}{3}(2+\rho)A_*(\alpha)H_\theta(\alpha)\sqrt{\theta}\rho^{-1}s^{-1/2} = A_*(\beta)\sqrt{\theta}H_\theta(\beta)\lambda^{-1/2} + \Big[-\frac{1}{6}\beta A_*(\beta)H_\theta(\beta) \\ \left(\beta H_\theta(\beta) - \beta^2\theta^{-1/2} - \sqrt{\theta}\right) - h_\theta(\beta)A_*(\beta)^2H_\theta(\beta)\sqrt{\theta}\Big]\lambda^{-1} + \mathcal{O}(\lambda^{-3/2}).\end{aligned} \tag{297}$$

Finally, substituting (294), (297) and $\mathcal{O}(s^{-3/2}) = \mathcal{O}(\lambda^{-3/2})$ into (292) yields the desired result. □

CHAPTER IV

FLUID MODELS FOR MANY-SERVER MARKOVIAN QUEUES IN CHANGING ENVIRONMENTS

4.1 Introduction

Fluid queues, or also called storage models, are queueing models where incoming work is continuous fluid, instead of discrete customers. Both stationary and transient behaviors of various fluid queues have been studied in the literature; for example, see [8, 33, 43] and the references therein.

One of the primary motivations for studying fluid queues is that they serve as first-order approximations for queueing systems with discrete customers and are often more tractable. In this chapter, we propose a fluid approximation for many-server Markovian queues where the arrival rate alternates between high and low values and study its stationary behavior.

Specifically, consider an $\mathrm{M}_t/\mathrm{M}/s{+}\mathrm{M}$ queue in a two-state random environment. The environment is modulated by a background process $\{L(t) : t \geq 0\}$, which alternates between a low-traffic state, say state 1 with arrival rate λ_1, and a high-traffic state, say state 2 with arrival rate λ_2. Every time that $L(t)$ reaches state i, it stays in that state for an exponentially distributed amount of time with mean $1/\nu_i$, $i = 1$ or 2, and then switches to the other state. We denote the service rate of each server by μ and thus when $L(t)$ is at state i, the traffic intensity is $\rho_i := \lambda_i/s\mu$. We assume that $\rho_1 < 1$ and $\rho_2 > 1$. Each customer has a patience time exponentially distributed with mean $1/\theta$. All random variables are independent of one another. We assume that $L(0) = 1$ or the system starts from a low-traffic state.

We propose approximating the queue-length process in the above system by a fluid queue described as follows. Consider a piecewise deterministic Markov process $\mathcal{M} := \{(X(t), L(t)) : t \geq 0\}$ with state space $[0, \infty) \times \{1, 2\}$. The evolution of $L(t)$ is the same as specified above. The evolution of $X(t)$ depends on $L(t)$ in the following way: When in state $(x, 1)$, $X(t)$ decreases at rate $r_1(x) := \mu(x \wedge s) + \theta(x - s)^+ - \lambda_1$; when in state $(x, 2)$,

$X(t)$ increases at rate $r_2(x) := \lambda_2 - \mu(x \wedge s) - \theta(x-s)^+$.

In terms of the connection between the approximating fluid queue and the corresponding system with discrete customers, the study by Choudhury et al. [13] is particularly relevant to our work. In [13], a fluid model is obtained as a law of large number limit with the rate at which the arrival rate (as well as other system parameters) changes converging to zero and the order of the number of servers held fixed. In contrast, we shall show that our fluid approximation, process $\mathcal{M}$, arises as the limit in a many-server asymptotic regime (similar to the model considered in [43]), where the number of servers grows to infinity and the arrival rate process is held fixed.

Our goal is to obtain the stationary distribution for the approximating fluid queue. To this end, we first note that state (x,l) is transient for all $x \in [0, \rho_1 s] \cup [x_m, \infty)$ and $l = 1$ or 2, where $\rho_1 s$ satisfies $r_1(\rho_1 s) = 0$ and x_m satisfies $r_2(x_m) = 0$ or

$$x_m := s + \frac{\lambda_2 - \mu s}{\theta} > s. \tag{298}$$

Because these transient states do not matter for the stationary behavior, throughout the remainder of this chapter, we study the process $\mathcal{M}$ with the state space being $S := (\rho_1 s, x_m) \times \{1,2\}$, instead of $[0,\infty) \times \{1,2\}$. For any state (x,l) in this bounded state space S, we have $r_l(x) > 0$. Therefore, our fluid queue falls into a class of on/off storage systems studied in [8], except for one minor difference on the state space.

Specifically, in [8] Boxma et al. study the stationary behavior of a two-dimensional Markov process, where the two dimensions, like in our model, correspond to the fluid level (or buffer content level) and the background, respectively. Also, their background switches between "on" and "off": When on (off), the fluid level increases (decreases) at some state-dependent, positive rate.

Their model is more general than ours in the following three aspects. First, their $r_i(x)$'s are not specifically defined but just some positive-valued functions with certain continuity properties, which are satisfied in our case. Second, the instantaneous rate at which their background state switches not only depends on the current background state but is influenced by the current fluid level. Hence, their switching rates are given by function $\lambda_i(x)$'s,

and in our model $\lambda_i(x) = \nu_i$ for each i. Third, their main result (Theorem 1 in [8]) covers two cases: one in which the stationary distribution has an atom at state $(x, 1)$ with x attaining the lower boundary value of the state space for the fluid level, and the other where no atom exists and the stationary distribution is given by a density function in the interior of the state space. Our model falls into the second case, as (x, l) with x equal to the lower boundary value $\rho_1 s$ is transient here.

The generality of the processes studied in [8] makes their analysis and results directly applicable to our setting. The only minor difference between our model and theirs is that their state space for the fluid level process is not bounded from above while ours $(\rho_1 s, x_m)$ is bounded from above. As will be seen in the next section, both our main result and the proof closely resemble those in [8].

Finally, we note that Scheinhardt et al. [45] also study the stationary (as well as transient) behavior of a very similar model by deriving the Kolmogorov equations. Like ours, their state space for the fluid level process is a bounded interval. In terms of the background proess, their model is the most general, allowing the background to switch among more than two states at rates depending on the current fluid level and background state. However, they assume that the fluid level rate functions $r_i(x)$'s to be strictly bounded away from 0 in the interior of the state space, and this assumption is not satisfied in our model because our $r_1(x)$ and $r_2(x)$ converge to 0 as $x \to \rho_1 s$ and $x \to x_m$, respectively.

In the next section, we shall investigate the stationary behavior of the Markov-modulated fluid queue process $\mathcal{M}$ using the same methodology as [8].

4.2 Main Result

4.2.1 Stationary distribution

In this section, we compute the stationary distribution for the Markov process $\mathcal{M}$. In what follows, any integral $\int_a^b f(x)dx$, with $a > b$, equals $-\int_b^a f(x)dx$.

Suppose a stationary distribution π exists and is given by a two-dimensional density function $\{(g_1(x), g_2(x)) : x \in (\rho_1 s, x_m)\}$, i.e., for any $[a, b] \subseteq (\rho_1 s, x_m)$, $\pi([a, b], i) = \int_a^b g_i(x)dx$,

$i = 1, 2$, and

$$\int_{\rho_1 s}^{x_m} g_1(x)dx + \int_{\rho_1 s}^{x_m} g_2(x)dx = 1. \tag{299}$$

Our main result is the explicit expression of the stationary density function.

Theorem 4.2.1. *For $i = 1, 2$,*

$$g_i(x) := \frac{C}{r_i(x)} \cdot \exp\left\{ -\int_s^x \left[\frac{\nu_2}{r_2(u)} - \frac{\nu_1}{r_1(u)} \right] du \right\} \tag{300}$$

$$= \begin{cases} \frac{C}{r_i(x)} \cdot \left(\frac{\lambda_2 - \mu x}{\lambda_2 - \mu s} \right)^{\frac{\nu_2}{\mu}} \cdot \left(\frac{\mu x - \lambda_1}{\mu s - \lambda_1} \right)^{\frac{\nu_1}{\mu}}, & \text{if } x \in (\rho_1 s, s] \\ \\ \frac{C}{r_i(x)} \cdot \left(\frac{\lambda_2 - \mu s - \theta(x-s)}{\lambda_2 - \mu s} \right)^{\frac{\nu_2}{\theta}} \cdot \left(\frac{\mu s + \theta(x-s) - \lambda_1}{\mu s - \lambda_1} \right)^{\frac{\nu_1}{\theta}}, & \text{if } x \in (s, x_m) \end{cases} \tag{301}$$

where

$$C := \frac{\nu_2}{\nu_1 + \nu_2} \cdot \frac{1}{C_1 + C_2}, \tag{302}$$

$$C_1 := (\lambda_2 - \mu s)^{-\frac{\nu_2}{\mu}} \cdot (\mu s - \lambda_1)^{-\frac{\nu_1}{\mu}} \cdot \int_{\rho_1 s}^{s} (\mu x - \lambda_1)^{\frac{\nu_1}{\mu} - 1} (\lambda_2 - \mu x)^{\frac{\nu_2}{\mu}} dx, \tag{303}$$

$$C_2 := (\lambda_2 - \mu s)^{-\frac{\nu_2}{\theta}} \cdot (\mu s - \lambda_1)^{-\frac{\nu_1}{\theta}} \cdot \int_{s}^{x_m} [\lambda_2 - \mu s - \theta(x-s)]^{\frac{\nu_2}{\theta}} [\mu s + \theta(x-s) - \lambda_1]^{\frac{\nu_1}{\theta} - 1} dx. \tag{304}$$

Note that (300) is exactly expression (7) in case (ii) of Theorem 1 in [8], if one replaces their ϵ with s, the switching rates $\lambda_i(x)$'s with ν_i's, the state space $(0, \infty)$ with $(\rho_1 s, x_m)$, and properly adjusts the normalizing constant (i.e., they assume $\int g_i(x)dx = 1$ for each i, instead of (299)) and background indices (i.e., they denote the on or high-traffic state by 0 instead of 2 here). Their proof also can be directly applied in our case after these minor modifications. For completeness we present the key steps of the proof.

Proof of Theorem 4.2.1. Fix an arbitrary $x \in (\rho_1 s, x_m)$. First, similar to equation (12) in [8], we have

$$g_1(x) r_1(x) = g_2(x) r_2(x). \tag{305}$$

This identity can be established by a rate-equality argument. Consider set $[x, x_m) \times \{1, 2\}$. The long-run average rate at which the process $\mathcal{M}$ leaves this set is given by $g_1(x)r_1(x)$ and the rate at which $\mathcal{M}$ enters this set $g_2(x)r_2(x)$. Then (305) immediately follows from equating thse two rates. We further define $h(x) := g_1(x)r_1(x) = g_2(x)r_2(x)$.

Next, similar to equation (16) in [8], we have

$$\int_x^{x_m} g_2(u)du \cdot \nu_2 = g_2(x)r_2(x) + \int_x^{x_m} g_1(u)du \cdot \nu_1, \tag{306}$$

which can be viewed as the rate-equality identity for set $[x, x_m) \times \{2\}$. Specifically, the rate at which the process $\mathcal{M}$ leaves this set is the left-hand side of (306) and the entering rate equals the right-hand side.

Substituting the definition of $h(x)$ into (306) yields

$$h(x) = \int_x^{x_m} h(u) \cdot \left[\frac{\nu_2}{r_2(u)} - \frac{\nu_1}{r_1(u)}\right] du. \tag{307}$$

We then solve (307) and find that

$$h(x) = h(s) \cdot \exp\left\{-\int_s^x \left[\frac{\nu_2}{r_2(u)} - \frac{\nu_1}{r_1(u)}\right] du\right\}. \tag{308}$$

Note that choosing the lower limit of integration in (308) as s is not essential, and one may set this lower limit to any $\epsilon \in (\rho_1 s, x_m)$, as done in case (ii) of Theorem 1 in [8].

From (308) and the definition of $h(x)$, we conclude that, for $i = 1, 2$,

$$g_i(x) = \frac{h(s)}{r_i(x)} \cdot \exp\left\{-\int_s^x \left[\frac{\nu_2}{r_2(u)} - \frac{\nu_1}{r_1(u)}\right] du\right\}. \tag{309}$$

Because the background switching rates are independent of the fluid level, we have

$$\int_{\rho_1 s}^{x_m} g_1(x)dx = \frac{\nu_2}{\nu_1 + \nu_2}. \tag{310}$$

We then substitute (309) into (310). Some simple calculations using the definitions of $r_i(x)$'s yield $h(s) = C$, where C is defined by (302) [1]. Replacing $h(s)$ with C in (309), we have (300) and eventually obtain (301) after some algebra. □

The existence and uniqueness of the stationary distribution follows from the counterpart of a set of five sufficient conditions specified in Section 4 of [8]. We next state these five conditions, all of which can be shown to hold by simple calculations. The sufficiency of this set of conditions follows from the same argument as provided in Section 4 of [8].

[1] One may also calculate $h(s)$ from $\int_{\rho_1 s}^{x_m} g_2(x)dx = \frac{\nu_1}{\nu_1+\nu_2}$ and obtain a different, yet equivalent expression for C.

Condition 1:

$$\int_{\rho_1 s}^{x_m} \frac{1}{r_1(x)} \cdot \exp\left\{-\int_s^x \left[\frac{\nu_2}{r_2(u)} - \frac{\nu_1}{r_1(u)}\right] du\right\} dx = C_1 + C_2 < \infty, \tag{311}$$

and

$$\int_{\rho_1 s}^{x_m} \frac{1}{r_2(x)} \cdot \exp\left\{-\int_s^x \left[\frac{\nu_2}{r_2(u)} - \frac{\nu_1}{r_1(u)}\right] du\right\} dx = \frac{\nu_1}{\nu_2} \cdot (C_1 + C_2) < \infty. \tag{312}$$

Condition 2:

$$\int_x^y \frac{1}{r_i(u)} du < \infty, \text{ for all } \rho_1 s < x < y < x_m \text{ and } i = 1, 2. \tag{313}$$

Condition 3:

$$\int_x^y \frac{\nu_i}{r_i(u)} du < \infty, \text{ for all } \rho_1 s < x < y < x_m \text{ and } i = 1, 2. \tag{314}$$

Condition 4:

$$\int_x^{x_m} \frac{1}{r_2(u)} du = \infty \quad \text{and} \quad \int_x^{x_m} \frac{\nu_2}{r_2(u)} du = \infty \quad \text{for some (hence all) } x \in (\rho_1 s, x_m). \tag{315}$$

Condition 5:

$$\int_{\rho_1 s}^y \frac{1}{r_1(u)} du = \infty \quad \text{and} \quad \int_{\rho_1 s}^y \frac{\nu_1}{r_1(u)} du = \infty \quad \text{for some (hence all) } y \in (\rho_1 s, x_m). \tag{316}$$

4.2.2 Slow-change asymptotics

Intuitively, if the rate at which the environment alternates is very low, the fluid level $X(t)$ can approach and stay close to the equilibrium value in each environment. Therefore, one expects that the stationary distribution defined by (301) should converge weakly to a discrete bimodal distribution taking on two possible values, $(\rho_1 s, 1)$ and $(x_m, 2)$, as ν_1 and ν_2 go to 0 in a proper manner. This intuition is made rigorous by our next result.

For any $\epsilon > 0$, let process $\mathcal{M}_\epsilon = \{X_\epsilon(t), L_\epsilon(t)\}$ be defined the same as $\mathcal{M}$ except that the environment alternating rates are given by $\nu_{i,\epsilon} := \epsilon\nu_i$ for some positive constant ν_i, $i = 1, 2$. Denote by π_ϵ the stationary distribution for $\mathcal{M}_\epsilon$ and by $\{g_{1,\epsilon}(x), g_{2,\epsilon}(x) : x \in (\rho_1 s, x_m)\}$ the corresponding density function, which can be specified by expression (301) with ν_i replaced by $\nu_{i,\epsilon}$, $i = 1, 2$. Let $\Rightarrow$ denote convergence in distribution.

Proposition 4.2.2. *$\pi_\epsilon \Rightarrow \pi_s$ as $\epsilon \to 0$, where π_s is a discrete probability distribution on $[0,\infty)\times[0,\infty)$, taking on the value $(\rho_1 s, 1)$ with probability $\nu_2/(\nu_1+\nu_2)$ and the value $(x_m, 2)$ with probability $\nu_1/(\nu_1+\nu_2)$.*

Proof. The cumulative distribution function (CDF) for π_s is

$$F_{\pi_s}(x,y) = \begin{cases} 0, & \text{if } (x,y) \in (-\infty, \rho_1 s) \times (-\infty, +\infty) \cup (-\infty, +\infty) \times (-\infty, 1) \\ \frac{\nu_2}{\nu_1+\nu_2}, & \text{if } (x,y) \in [\rho_1 s, x_m) \times [1, \infty) \cup [x_m, \infty) \times [1, 2) \\ 1, & \text{otherwise} \end{cases}. \tag{317}$$

The set of continuity points of $F_{\pi_s}(\cdot,\cdot)$ is $\mathcal{C} := \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$, where

$$\mathcal{C}_1 = (-\infty, \rho_1 s) \times (-\infty, +\infty) \cup (-\infty, +\infty) \times (-\infty, 1), \tag{318}$$

$$\mathcal{C}_2 = (\rho_1 s, x_m) \times (1, \infty) \cup [x_m, \infty) \times (1, 2), \tag{319}$$

and

$$\mathcal{C}_3 = (x_m, \infty) \times (2, \infty). \tag{320}$$

Denote by $F_{\pi_\epsilon}(\cdot,\cdot)$ the CDF for π_ϵ. It is easy to see that for any $(x,y) \in \mathcal{C}_1$, $F_{\pi_\epsilon}(x,y) = F_{\pi_s}(x,y) = 0$ and for any $(x,y) \in \mathcal{C}_3$, $F_{\pi_\epsilon}(x,y) = F_{\pi_s}(x,y) = 1$. It follows from (301) that for any $x \in (\rho_1 s, s]$,

$$\begin{aligned} F_{\pi_\epsilon}(x,1) &= \int_{\rho_1 s}^{x} g_{1,\epsilon}(u) du \\ &= \frac{\nu_{2,\epsilon}}{\nu_{1,\epsilon}+\nu_{2,\epsilon}} \cdot \frac{C_{1,\epsilon}(x)}{C_{1,\epsilon}(s)+C_{2,\epsilon}}, \end{aligned} \tag{321}$$

where

$$C_{1,\epsilon}(x) := (\lambda_2 - \mu s)^{-\frac{\nu_{2,\epsilon}}{\mu}} (\mu s - \lambda_1)^{-\frac{\nu_{1,\epsilon}}{\mu}} \int_{\rho_1 s}^{x} (\mu u - \lambda_1)^{\frac{\nu_{1,\epsilon}}{\mu}-1} (\lambda_2 - \mu u)^{\frac{\nu_{2,\epsilon}}{\mu}} du, \forall x \in (\rho_1 s, s], \tag{322}$$

and

$$C_{2,\epsilon} := (\lambda_2 - \mu s)^{-\frac{\nu_{2,\epsilon}}{\theta}} \cdot (\mu s - \lambda_1)^{-\frac{\nu_{1,\epsilon}}{\theta}} \cdot \int_{s}^{x_m} [\lambda_2 - \mu s - \theta(x-s)]^{\frac{\nu_{2,\epsilon}}{\theta}} [\mu s + \theta(x-s) - \lambda_1]^{\frac{\nu_{1,\epsilon}}{\theta}-1} dx. \tag{323}$$

Using the Dominated Convergence Theorem, we obtain that

$$\lim_{\epsilon\to 0} C_{2,\epsilon} = \int_s^{x_m} [\mu s + \theta(x-s) - \lambda_1]^{-1} dx = \frac{1}{\theta} \ln \frac{\lambda_2 - \lambda_1}{\mu s - \lambda_1}. \tag{324}$$

Some simple analysis further shows that $\lim_{\epsilon\to 0} C_{1,\epsilon}(x) = \infty$ and $C_{1,\epsilon}(x) \sim C_{1,\epsilon}(s)$ as $\epsilon \to 0$, which together with (321) and (324) yield that

$$\lim_{\epsilon\to 0} F_{\pi_\epsilon}(x,1) = \frac{\nu_2}{\nu_1+\nu_2}, \text{ for any } x \in (\rho_1 s, s]. \tag{325}$$

Because for any $x > s$, $F_{\pi_\epsilon}(x,1) \in [F_{\pi_\epsilon}(s,1), \frac{\nu_2}{\nu_1+\nu_2}]$, it follows that (325) actually holds for all $x \in (\rho_1 s, \infty)$. Combining this with the fact that $F_{\pi_\epsilon}(x,y) \in [F_{\pi_\epsilon}(x,1), \frac{\nu_2}{\nu_1+\nu_2}]$ for all $(x,y) \in \mathcal{C}_2$, we find that

$$\lim_{\epsilon\to 0} F_{\pi_\epsilon}(x,y) = \frac{\nu_2}{\nu_1+\nu_2} = F_{\pi_s}(x,y), \text{ for any } (x,y) \in \mathcal{C}_2. \tag{326}$$

In conclusion, $\lim_{\epsilon\to 0} F_{\pi_\epsilon}(x,y) = F_{\pi_s}(x,y)$ for any $(x,y) \in \mathcal{C}$ and hence $\pi_\epsilon \Rightarrow \pi_s$ as $\epsilon \to 0$.

□

4.2.3 Many-server fluid limit

Finally, we establish a many-server limit theorem for the Markovian queueing system in an environment modulated by $\{L(t) : t \geq 0\}$.

Consider a sequence of multi-server queues indexed by the number of servers, or in the Nth system the number of servers $s^N := N$. When $L(t) = i$, the customer arrival process in the Nth system is a Poisson process with rate λ_i^N, which satisfies

$$\lim_{N\to\infty} \frac{\lambda_i^N}{N} = \frac{\lambda_i}{s}, \ i = 1, 2. \tag{327}$$

Each customer requires a service time exponentially distributed with mean $1/\mu$ and has a patience time exponentially distributed with mean $1/\theta$. Both customer service times and patience times are independent of N or i. We denote the traffic intensity in the Nth system when $L(t) = i$ by $\rho_i^N := \lambda_i^N / N\mu$ and assume that

$$\lim_{N\to\infty} \sqrt{N}(\rho_i - \rho_i^N) = \beta_i, \tag{328}$$

for some $-\infty < \beta_i < \infty$, $i = 1, 2$.

Denote by $Q^N(t)$ the number of customers in the Nth system at time t and let $\bar{Q}^N(t) := N^{-1}Q^N(t)$. Let $D[0,\infty)$ denote the space of real-valued functions that are right-continuous on $[0,\infty)$ and have left limits everywhere in $(0,\infty)$, endowed with the usual Skorohod topology.

Theorem 4.2.3. *If there exists a random variable $\bar{Q}_0$ such that $\bar{Q}^N(0) \Rightarrow \bar{Q}_0$ as $N \to \infty$, then*

$$\bar{Q}^N \Rightarrow \bar{Q} \quad in \quad D[0,\infty) \quad as \quad N \to \infty, \tag{329}$$

where $\bar{Q} := \{\bar{Q}(t) : t \geq 0\}$ is such that $\bar{Q}(0) = \bar{Q}_0$ and $\{s\bar{Q}(t) : t \geq 0\}$ has the same evolution as the first dimension of the process $\mathcal{M}$, i.e., $\{X(t) : t \geq 0\}$.

Proof. Recall that we assume $L(0) = 1$. Let us denote by $Y_{l,i}$ the length of the ith low-traffic time interval and by $Y_{h,i}$ that of the ith high-traffic time interval, for $i \geq 1$. Further define $T_{h,0} := 0$,

$$T_{h,j} := \sum_{i=1}^{j} (Y_{l,i} + Y_{h,i}),\ j \geq 1, \tag{330}$$

and

$$T_{l,j} := T_{h,j-1} + Y_{l,j}, j \geq 1. \tag{331}$$

So we have $L(t) = 1$, $t \in [T_{h,j-1}, T_{l,j})$, and $L(t) = 2$, $t \in [T_{l,j}, T_{h,j})$, for all $j \geq 1$. Or in words, $[T_{h,j-1}, T_{l,j})$ represents the jth low-traffic time interval and $[T_{l,j}, T_{h,j})$ the jth high-traffic time interval.

Define the process $\bar{Q}$ as follows: for $t \in (T_{h,j-1}, T_{l,j}]$, $\forall j \geq 1$,

$$\bar{Q}(t) = \bar{Q}(T_{h,j-1}) + \frac{\lambda_1}{s} \cdot (t - T_{h,j-1}) - \mu \int_{T_{h,j-1}}^{t} (\bar{Q}(u) \wedge 1) du - \theta \int_{T_{h,j-1}}^{t} (\bar{Q}(u) - 1)^+ du, \tag{332}$$

and for $t \in (T_{l,j}, T_{h,j}]$, $\forall j \geq 1$,

$$\bar{Q}(t) = \bar{Q}(T_{l,j}) + \frac{\lambda_2}{s} \cdot (t - T_{l,j}) - \mu \int_{T_{l,j}}^{t} (\bar{Q}(u) \wedge 1) du - \theta \int_{T_{l,j}}^{t} (\bar{Q}(u) - 1)^+ du. \tag{333}$$

It is easy to see that $\{s\bar{Q}(t) : t \geq 0\}$ has the same evolution as the first dimension of the process $\mathcal{M}$, or $\{X(t) : t \geq 0\}$, if they start from the same initial state.

To establish weak convergence, we first fix an arbitrary realization of the environment process L (or equivalantly a realization of $\{(Y_{l,i}, Y_{h,i}) : i \geq 1\}$). Because $\bar{Q}^N(0) \Rightarrow \bar{Q}(0)$ as

$N \to \infty$, it follows from the argument in Section EC.2.3 of [43] (specifically, the proof from equation (EC.22) to equation (EC.29) in the e-Companion to [43]) that the processes $\bar{Q}^N$ restricted to the time interval $[0, T_{l,1})$ converge weakly to the process $\bar{Q}$ restricted to the time interval $[0, T_{l,1})$ in $D[0, \infty)$, and hence $\bar{Q}^N(T_{l,1}-) \Rightarrow \bar{Q}(T_{l,1}-)$. Then the same argument yields that the processes $\bar{Q}^N$ restricted to the time interval $[T_{l,1}, T_{h,1})$ converge weakly to the process $\bar{Q}$ restricted to the time interval $[T_{l,1}, T_{h,1})$ in $D[0, \infty)$ and therefore $\bar{Q}^N(T_{h,1}-) \Rightarrow \bar{Q}(T_{h,1}-)$. Continuing this recursive argument leads to the weak convergence on every low-traffic and high-traffic time interval. This allows us to conclude that conditioning on L, the processes $\bar{Q}^N$ converge to the process $\bar{Q}$ in $D[0, \infty)$ and hence without conditioning $\bar{Q}^N \Rightarrow \bar{Q}$ in $D[0, \infty)$.

□

REFERENCES

[1] Abramowitz, M. and Stegun, I. A., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover, tenth printing, december 1972, with corrections ed., 1964.

[2] Asmussen, S., *Applied Probability and Queues (2nd edition)*. Springer, 2003.

[3] Bachmat, E. and Sarfati, H., "Analysis of SITA policies," *Perform. Eval.*, vol. 67, no. 2, pp. 102–120, 2010.

[4] Bassamboo, A. and Randhawa, R. S., "On the accuracy of fluid models for capacity planning in queueing systems with impatient customers," *Preprint*, 2009.

[5] Blanchet, J. and Glynn, P., "Complete corrected diffusion approximations for the maximum of a random walk," *Ann. Appl. Probab.*, vol. 16, no. 2, pp. 951–983, 2006.

[6] Blanchet, J., Glynn, P., and Lam, H., "Rare event simulation for a slotted time $M/G/s$ model," *Queueing Syst. Theory Appl.*, vol. 63, pp. 33–57, 2009.

[7] Borst, S., Mandelbaum, A., and Reiman, M., "Dimensioning large call centers," *Oper. Res.*, vol. 52, pp. 17–34, 2004.

[8] Boxma, O., Kaspi, H., Kella, O., and Perry, D., "On/off storage systems with state-dependent input, output, and switching rates," *Probab. Eng. Inf. Sci.*, vol. 19, no. 1, pp. 1–14, 2005.

[9] Bradley, J. and Glynn, P., "Managing capacity and inventory jointly in manufacturing systems," *Manage. Sci.*, vol. 48, no. 2, pp. 273–288, 2002.

[10] Brigandi, A. J., Dargon, D. R., Sheenan, M. J., and III, T. S., "AT&T's call processing simulator (CAPS) operational design for inbound call centers.," *Interfaces*, vol. 24, no. 1, pp. 6 – 28, 1994.

[11] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L., "Statistical analysis of a telephone call center: A queueing-science perspective," *J. Am. Stat. Assoc.*, vol. 100, pp. 36–50, 2005.

[12] Cardellini, V., Casalicchio, E., Colajanni, M., and Yu, P. S., "The state of the art in locally distributed web-server systems," *ACM Computing Surveys*, vol. 34, no. 2, pp. 263–311, 2002.

[13] Choudhury, G. L., Mandelbaum, A., Reiman, M. I., and Whitt, W., "Fluid and diffusion limits for queues in slowly changing environments," *Stoch. Mod*, vol. 13, pp. 121–146, 1997.

[14] Ciardo, G., Riska, A., and Smirni, E., "Equiload: a load balancing policy for clustered web servers," *Performance Evaluation*, vol. 46, pp. 46–101, 2001.

[15] Dai, J. G., He, S., and Tezcan, T., "Many-server diffusion limits for G/Ph/n + GI queues," *Preprint*, 2009.

[16] Embrechts, P., Klüppelberg, C., and Mikosch, T., *Modelling Extremal Events: for Insurance and Finance (Stochastic Modelling and Applied Probability)*. Springer, 2008.

[17] Finkenstadt, B. and Rootzen, H., *Extreme Values in Finance, Telecommunications, and the Environment*. CRC Press LLC, 2004.

[18] Gamarnik, D. and Momčilović, P., "Steady-state analysis of a multiserver queue in the Halfin-Whitt regime," *Adv. in Appl. Probab.*, vol. 40, no. 2, pp. 548–577, 2008.

[19] Ganesh, A., O'Connell, N., and Wischik, D., *Big Queues*. Springer, 2004.

[20] Gans, N., Koole, G., and Mandelbaum, A., "Telephone call centers: Tutorial, review, and research prospects," *Manufacturing Service Oper. Management*, pp. 79–141, 2003.

[21] Garnett, O., Mandelbaum, A., and Reiman, M., "Designing a call center with impatient customers," *Manufacturing Service Oper. Management*, vol. 4, pp. 208–227, 2002.

[22] Graham, R. L., Grötschel, M., and Lovász, L., *Handbook of Combinatorics*. The MIT Press, 2003.

[23] Gross, D. and Harris, C. M., *Fundamentals of Queueing Theory (3rd Edition)*. Wiley-Interscience, 1998.

[24] Halfin, S. and Whitt, W., "Heavy-traffic limits for queues with many exponential servers," *Oper. Res.*, vol. 29, pp. 567–588, 1981.

[25] Harchol-Balter, M., "Queueing disciplines," in *Wiley Encyclopedia Of Operations Research and Management Science*, 2009.

[26] Harchol-balter, M., Crovella, M. E., and Murta, C. D., "On choosing a task assignment policy for a distributed server system," *Journal of Parallel and Distributed Computing*, vol. 59, pp. 204–228, 1999.

[27] Harchol-Balter, M., Scheller-Wolf, A., and Young, A., "Surprising results on task assignment in server farms with high-variability workloads," *Proceedings of ACM SIGMETRICS 2009 Conference on Measurement and Modeling of Computer Systems*, 2009.

[28] Harchol-Balter, M., Scheller-Wolf, A., and Young, A., "Why segregating short jobs from long jobs under high variability is not always a win," *Allerton Conference, Urbana, Champaign*, pp. 102–120, 2009.

[29] Janssen, A., van Leeuwaarden, J., and Zwart, B., "Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula," *Adv. in Appl. Probab.*, vol. 40, no. 1, pp. 122–143, 2008.

[30] Janssen, A., van Leeuwaarden, J., and Zwart, B., "Refining square root safety staffing by expanding Erlang C," *To appear in Oper. Res.*, 2008.

[31] Jelenković, P., Mandelbaum, A., and Momčilović, P., "Heavy traffic limits for queues with many deterministic servers," *Queueing Syst. Theory Appl.*, vol. 47, no. 1/2, pp. 53–69, 2004.

[32] Kang, W. and Ramanan, K., "Fluid limits of many-servers queues with reneging," *Preprint*, 2008.

[33] Kella, O. and Whitt, W., "A storage model with a two-state random environment," *Oper. Res.*, vol. 40, pp. 257–262, 1991.

[34] Kingman, J., "The first Erlang century–and the next," *Queueing Syst. Theory Appl.*, vol. 63, pp. 3–12, 2009.

[35] Kumar, S. and Randhawa, R. S., "Exploiting market size in service systems," *To appear in Manufacturing Service Oper. Management*, 2009.

[36] Liu, H. and Wee, S., "Web server farm in the cloud: Performance evaluation and dynamic architecture," in *Cloud Computing*, vol. 5931, pp. 369–380, Springer Berlin / Heidelberg, 2009.

[37] López-Ortiz, A., "Valiant load balancing, capacity provisioning and resilient backbone design," in *Combinatorial and Algorithmic Aspects of Networking*, vol. 4852, pp. 3–12, Springer Berlin / Heidelberg, 2007.

[38] Mandelbaum, A. and Momcilovic, P., "Queues with many servers and impatient customers," *Preprint*, 2009.

[39] Mandelbaum, A. and Zeltyn, S., "Service engineering in action: The Palm/Erlang-A queue, with applications to call centers," in *Advances in Services Innovations*, pp. 17–48, Springer-Verlag, 2007.

[40] Mandelbaum, A. and Zeltyn, S., "Staffing many-server queues with impatient customers: constraint satisfaction in call centers," *Under revision for Oper. Res.*, 2008.

[41] Nuyens, M., Wierman, A., and Zwart, B., "Preventing large sojourn times using smart scheduling," *Oper. Res.*, vol. 56, no. 1, pp. 88–101, 2008.

[42] Pang, G. and Whitt, W., "Service interruptions in large-scale service systems," *Management Sci.*, vol. 55, no. 9, pp. 1499–1512, 2009.

[43] Pang, G. and Whitt, W., "Service interruptions in large-scale service systems," *Manage. Sci.*, vol. 55, pp. 1499–1512, 2009.

[44] Reed, J. and Zhang, B., "Managing capacity and inventory jointly for large-scale manufacturing systems: square-root rules," *Working paper*, 2011.

[45] Scheinhardt, W., Foreest, N., and Mandjes, M., "Continuous feedback fluid queues," *Oper. Res. Lett.*, vol. 33, pp. 551–559, 2005.

[46] Schroeder, B. and Harchol-Balter, M., "Evaluation of task assignment policies for supercomputing servers: The case for load unbalancing and fairness," *Cluster Computing*, vol. 7, no. 2, pp. 151–161, 2004.

[47] Siegmund, D., "Corrected diffusion approximations in certain random walk problems," *Adv. Appl. Prob.*, vol. 11, no. 4, pp. 701–719, 1979.

[48] Smith, D. R. and Whitt, W., "Resource sharing for efficiency in traffic systems," *the Bell System Technical Journal*, vol. 60, no. 1, pp. 39–55, 1981.

[49] Stanley, R. P., *Enumerative Combinatorics, Volume 1 (2nd edition)*. Cambridge University Press, 2000.

[50] Tezcan, T. and Dai, J., "Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic," *To appear in Oper. Res.*, 2008.

[51] Tijms, H. C., *Stochastic Models: An Algorithmic Approach*. John Wiley & Sons, 1995.

[52] Tijms, H. C., *A First Course in Stochastic Models*. John Wiley & Sons, 2003.

[53] Tijms, H. C., Hoorn, M. H. V., and Federgruen, A., "Approximations for the steady-state probabilities in the M/G/c queue," *Adv. in Appl. Probab.*, vol. 13, no. 1, pp. 186–206, 1981.

[54] Wang, C. and Wolff, R. W., "The M/G/c queue in light traffic," *Queueing Syst. Theory Appl.*, vol. 29, no. 1, pp. 17–34, 1998.

[55] Whitt, W., "The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution," *Queueing Syst. Theory Appl.*, vol. 36, pp. 71–87, 2000.

[56] Whitt, W., "The Erlang B and C formulas: Problems and solutions," *Class notes*, 2002.

[57] Whitt, W., "Two fluid approximations for multi-server queues with abandonments," *Oper. Res. Lett.*, vol. 33, pp. 363–372, 2005.

[58] Whitt, W., "Fluid models for multiserver queues with abandonments," *Oper. Res.*, vol. 54, no. 1, pp. 37–54, 2006.

[59] Whitt, W., "Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters," *Oper. Res.*, vol. 54, no. 2, pp. 247–260, 2006.

[60] Wierman, A. and Zwart, B., "Is tail-optimal scheduling possible?," *Preprint*, 2010.

[61] Yao, D. D., "Refining the diffusion approximation for the $M/G/m$ queue," *Oper. Res.*, vol. 33, no. 6, pp. 1266–1277, 1985.

[62] Zeltyn, S. and Mandelbaum, A., "Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue," *Queueing Syst. Theory Appl.*, vol. 51, pp. 361–402, 2005.

[63] Zhang, J., "Fluid models of many-server queues with abandonment," *Preprint*, 09 2009.